VŠB – Technická univerzita Ostrava
Fakulta elektrotechniky a informatiky
Katedra informatiky

Analýza chování uživatelů adaptivního webu

Analysis of User Behavior of Adaptive Web

2013 Bc. Martin Hruzík

VŠB - Technická univerzita Ostrava
Fakulta elektrotechniky a informatiky
Katedra informatiky

Zadání diplomové práce

Student: **Bc. Martin Hruzík**

Studijní program: N2647 Informační a komunikační technologie

Studijní obor: 2612T025 Informatika a výpočetní technika

Téma: Analýza chování uživatelů adaptivního webu
Analysis of User Behavior of Adaptive Web

Zásady pro vypracování:

Chování uživatele webu lze analyzovat statisticky, v případě existence sémantické informace, jako je například informace o formuláři, dotazníku, rozhodovacím bloku, nebo obsahu pro čtení, příkladu, potřebě aktivní odpovědi, lze významně zkvalitnit analýzu a upravit model uživatele. Ten může být sdílen a může být využitelný pro personalizované webové systémy ve stejné, ale i jiné aplikační doméně.
Na VŠB-TU Ostrava byl vytvořen experimentální adaptivní systém XAPOS, podrobnější informace na http://arg.vsb.cz/xapos/
Cílem práce je formulace hypotézy o zkvalitnění navigace na základě znalosti významu a kategorie obsahu (vycházející ze zjištění diplomanta, po konzultaci s vedoucím), realizace případové studie, její vyhodnocení a zpětná vazba k adaptovanému obsahu ve vzdělávací doméně.

Seznam doporučené odborné literatury:

[1] Bing Liu. Web DataMining; Exploring Hyperlinks, Contents, and Usage Data. Chapter Link Analysis, pages 237-272. Corrected 2nd printing 2008 ISBN-10 3-540-37881-2 Springer Berlin Heidelberg New York, 2007.
[2] Peter Brusilovsky, Alfred Kobsa, Wolfgang Nejdl (Eds.). The Adaptive Web; Methods and Strategies of Web Personalization. LNCS4321, ISBN-10 3-540-72078-2 Springer, 2007
[3] Tvarožek, M., Barla, M., Bieliková, M. (2007). Personalized Presentation in Web-Based Information Systems. In J. Van Leeuwen, G. F. Italiano, W. van der Hoek, H. Sack, C. Meinel, F. Plášil (Ed.), SOFSEM 2007: Proceedings of the 33rd Conference on Current Trends in Theory and Practice of Computer Science. LNCS 4362, pp. 796-807. Harrachov, Czech Republic: Springer-Verlag, Berlin Heidelberg.
[4] Andrejko, A., Barla, M., Bieliková, M., Tvarožek, M. (2006). Softvérové nástroje pre získavanie charakteristík používatela. In P. Vojtáš & T. Skopal (Ed.), Proceedings of DATAKON '06, (pp. 139-148). Brno, Czech Republic (in Slovak).
[5] A. K. Jain, M. N. Murty, P. J. Flynn. Data clustering: a review. ACM Computing Surveys (CSUR), Volume 31 , Issue 3 (September 1999), ISSN:0360-0300, Pages: 264 - 323, ACM 1999
[6] De Bra P., Houben G., Wu H. AHAM: A Dexter-based Reference Model for Adaptive Hypermedia. In Proceedings of the ACM Conference on Hypertext and Hypermedia, Darmstadt, Germany, 1999, p. 221-239.
[7] Šaloun P., Velart Z., Concept Space Rating for Personalization of Learning Materials Based on Relations. In. Proc. of SMAP'09, 4th International Workshop on Semantic Media Adaptation and Personalization, CS IEEE 2009, pp. 67-72 (2009).
[8] Nástroj RaVis. Online http://code.google.com/p/birdeye/wiki/RaVis
[9] PSPP, online http://www.gnu.org/software/pspp/

Formální náležitosti a rozsah diplomové práce stanoví pokyny pro vypracování zveřejněné na webových stránkách fakulty.

Vedoucí diplomové práce: **doc. RNDr. Petr Šaloun, Ph.D.**

Datum zadání: 16.11.2012

Datum odevzdání: 07.05.2013

doc. Dr. Ing. Eduard Sojka
vedoucí katedry

prof. RNDr. Václav Snášel, CSc.
děkan fakulty

Souhlasím se zveřejněním této diplomové práce dle požadavků čl. 26, odst. 9 *Studijního a zkušebního řádu pro studium v magisterských programech VŠB-TU Ostrava.*

V Ostravě 31.7.2013 .....

Prohlašuji, že jsem tuto diplomovou práci vypracoval samostatně. Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

V Ostravě 31.7.2013 .....

Rád bych poděkoval doc. RNDr. Petru Šalounovi, Ph.D. za odbornou konzultaci, trpělivost a pomoc při psaní této diplomové práce. Za podporu bych chtěl poděkovat své rodině, přátelům a svému zaměstnavateli.

Abstrakt

V posledních letech dochází k silnému rozvoji informačních technologií a pokračuje nárust počtu uživatelů internetu. Díky rychle přibývajícímu množství dat ale nastává problém vhodného výběru a doručení kvalitních relevantních informací koncovým uživatelům. Tímto problémem se zabývají hlavně vyhledávače, ale rovněž se jej snaží řešit samotné weby. Snaží se chápat jednotlivé uživatele a adaptovat obsah dle jejich potřeb. Jednou z klíčových oblastí je samotná navigace na webu, která uživatele směřuje k danému cíli. Velmi diskutovanou oblastí je analýza sentimentu, která se zabývá dolováním emocí z textu. Takto získaná data mají často nesmírnou cenu. Pro česky psané texty se však zatím jedná o velmi málo probádanou oblast, kterou se tato práce snaží odkrýt.

Klíčová slova: adaptivní web, sémantický web, analýza sentimentu, dolování názoru, vizualizace dat, hodnocení obsahu

Abstract

In recent years, there has been a significant development of information technology, and the number of the Internet users keeps increasing. However, due to the fact that the amount of the available data is growing rapidly, the problem of delivering relevant and quality data to end users occurs. Search engines are primarily concerned with this issue, but web pages themselves are trying to address this as well. Web pages are trying to understand individual users and adapt the contents accordingly. One of the key areas is the web navigation, which is what steers user to a particular target. Sentiment analysis is a much discussed topic, which is concerned with capturing emotion from a given text. This information is often of significant value. As far as Czech texts are concerned, this area is largely unexplored, and this work is attempting to change it.

Keywords: adaptive web, semantic web, sentiment analysis, opinion mining, data visualization, content evaluation

Seznam použitých zkratek a symbolů

GUI	–	Graphical User Interface, grafické uživatelské rozhraní
HTML	–	HyperText Markup Language, značkovací jazyk pro hypertext
MVC	–	Model-view-controler, model-pohled-řadič
LDAP	–	Lightweight Directory Access Protocol
NLP	–	Natural Language Processing, zpracování přirozeného jazyka
SQL	–	Structured Query Language, strukturovaný dotazovací jazyk

Obsah

Seznam tabulek

Seznam obrázků

Seznam výpisů zdrojového kódu

1 Úvod

S postupným vývojem internetu se velice rychle rozvíjejí možnosti využití. Internet se s postupem času stal součástí každodenního života a řada lidí si život bez něj již nedokáže představit. Zatím co před několika lety šlo o vymoženost, dnes je internet většinou považován za samozřejmost. Přístup k internetu byl dokonce navržen na přidání do základní listiny lidských práv [1].

S velmi rychlým nárůstem objemu informací napříč celým internetem ale začal vznikat problém plynoucí z velkého množství dat. Začátkem třetího tisíciletí byla vyslovena myšlenka sémantického webu. Jejím autorem je Tim Bernes-Lee, který upozornil na to, že se z webu stala směs různých stránek a najít relevantní informace je stále těžší a těžší. Cílem sémantického webu je přidělit informacím jejich význam a popsat je. Výsledkem je možnost automatizovaného zpracování a přiblížení relevantnějších informací jednotlivým uživatelům.

Velkým zásahem do světa webu se stal rok 2004, kdy byl představen standard Web 2.0 [10]. Nešlo o změnu technických norem, ale o změnu v chápání webu jako takového. Stránky začaly plnit interaktivní prvky a uživatelé se začali na tvorbě obsahu aktivně podílet. Uživatel přestal být jen konzument obsahu, ale začal být jeho aktivním spolutvůrcem. Weby začaly získávat spoustu informací o svých návštěvnících či zákaznících a na řadu přišla otázka, jak s těmito informacemi naložit.

Adaptivní web začal data o jednotlivých uživatelích využívat. Web už tak nepřistupoval ke všem uživatelům stejně, ale využíval svých uložených informací k adaptaci a personalizaci. Nejznámějším příkladem jsou internetové vyhledávače, které přizpůsobují výsledky vyhledávání našim potřebám, ale také využívají personalizace k zobrazování reklam. S adaptací navigace a zobrazovaného obsahu se zase setkáváme u řady zpravodajských webů.

Weby se snaží získávat o uživateli co nejvíce informací a hledají možnosti, jak tyto cenné informace využít k úpravě nabídky, navigace a hlavně doručení relevantního obsahu na míru uživateli. Zdrojem informací o uživateli může být zaznamenávání jeho průchodu webem, hlasování v anketě, ohodnocení obsahu nebo například udělení komentáře. Všechny tyto informace mohou pomoci ke zkvalitňování webu o jeho obsahu.

Tato diplomová práce je zasazena do prostředí adaptivního webu, kde informace o uživateli má skutečný význam. Práce se zaměřuje na chování uživatelů a vyhodno-

[1] http://www2.ohchr.org/english/bodies/hrcouncil/docs/17session/A.HRC.17.27_en.pdf

cení dat, která o sobě zanechají. Jedná se o zaznamenané přechody mezi stránkami, komentáře nebo přidané anotace a jejich hodnocení. Hlavní část práce se zaměřuje na analýzu sentimentu komentářů a hledání jejich významů. Rozpoznávání emocí z textů je aktivně diskutované téma. Hledání sentimentu v česky psaných textech je stále oblastí, kde je mnoho prostoru k experimentům, výzkumu a zvýšení úspěšnosti analýzy sentimentu.

Práce je rozdělena do osmi kapitol. Vývoj webu a stručný popis prostředí sémantického a adaptivního webu obsahuje kapitola 2. V kapitole 3 je popsán adaptivní e-learningový systém XAPOS a experiment, který v něm byl navržen, naplánován a proveden. Obsažen je přehled získaných dat se srovnáním výsledků studentů s hodnocením v předmětu.

Kapitola 4 se věnuje navigaci, jejímu současnému stavu a možnostem úpravy na základě analýzy chování uživatelů webu. Popsán je také proces vizualizace průchodu systémem a analýza chování uživatele.

V kapitole 5 je podrobně popsána analýza sentimentu a její možnosti. Získávání emocí z textu je věnována většina této práce a je představen i vlastní přístup. Ten je navržen na základě různých metod používaných při analýze sentimentu v anglických textech, ale rovněž jsou využity vlastní prvky a také vlastní jazykový cit, který čeština při analýze vyžaduje. Kapitola představuje hlavní přínos této diplomové práce a její jádro a experiment jsou shrnuty do výzkumného článku uvedeného v plném znění v příloze A.

V poslední kapitole 6 jsou prezentovány výsledky z analýzy chování uživatelů a analýzy sentimentu. Jsou obsaženy také návrhy pro další práci.

V příloze A je obsažen článek prezentující výsledky výzkumu a experimentu této práce na téma analýzy sentimentu. Článek byl přijat na konferenci "Interdisciplinary Symposium on Complex Systems - ISCS2013"[2], konanou ve dnech 10. - 13. září 2013 v Praze, kde bude prezentován odborné veřejnosti. V přiložené podobě bude článek publikován v knize vydané nakladatelstvím Springer.

[2]https://sites.google.com/site/complexsystems2013/home

2 Vývoj webu

Svět internetu a webu prodělal za poslední desítku let obrovský pokrok. Statické webové stránky od roku 2004 začaly nahrazovat stránky plné dynamických prvků. Tehdejší uvedení Web 2.0 bylo skutečnou internetovou revolucí. Nejednalo se však o změnu HTML standardu, ale o etapu vývoje webu, kdy začaly být statické weby nahrazovány dynamickými a na tvorbě obsahu se začali podílet samotní uživatelé.

Hlavním cílem zavedení Web 2.0 byla změna v chápání webu jako takového. Uživatel již neměl být jen pasivním návštěvníkem prohlížejícím si obsah. Naopak má být do tvorby obsahu webu přímo vtažen. Pomoci mu v tom mají mimo jiné četné interaktivní prvky. Svět informačních technologií ale od roku 2004 výrazně pokročil a při pohledu na dnešní weby se již často jedná o propracované webové aplikace, které mohou být dokonce plnohodnotnou náhradou těch desktopových. Kam se Web 2.0 za prvních pět let své existence dostal je popsáno v [10]. Na internet se v posledních letech přesunulo mnoho služeb a ve velké míře rovněž mezilidská komunikace. Možnosti internetu ještě navýšil masivní rozvoj sociálních sítí [28].

Již dlouhou dobu se objevují četné zmínky o definování Web 3.0. Neoficiální popis uvádí, že by se měl soustředit na individualitu uživatelů a jejich potřeby. Stejně tak na využití digitálních médií. V zásadě se jedná o web sémantický, který ještě více zaměřen na personalizaci a adaptaci obsahu. Přináší také další možnosti interaktivity a využití umělé inteligence. Zdá se tedy, že vývoj internetu je velmi dynamický a jeho možnosti ještě zdaleka nejsou vyčerpány.

2.1 Sémantický web

S rapidně rostoucím světem internetu a zejména nárůstem obsahu informací v něm obsažených ale přichází nutnost rozeznávat kvalitu obsahu a doručovat uživateli taková data, která on sám chce a očekává. Hledání relevantních informací v tak nepřeberném množství dat ale není nic jednoduchého.

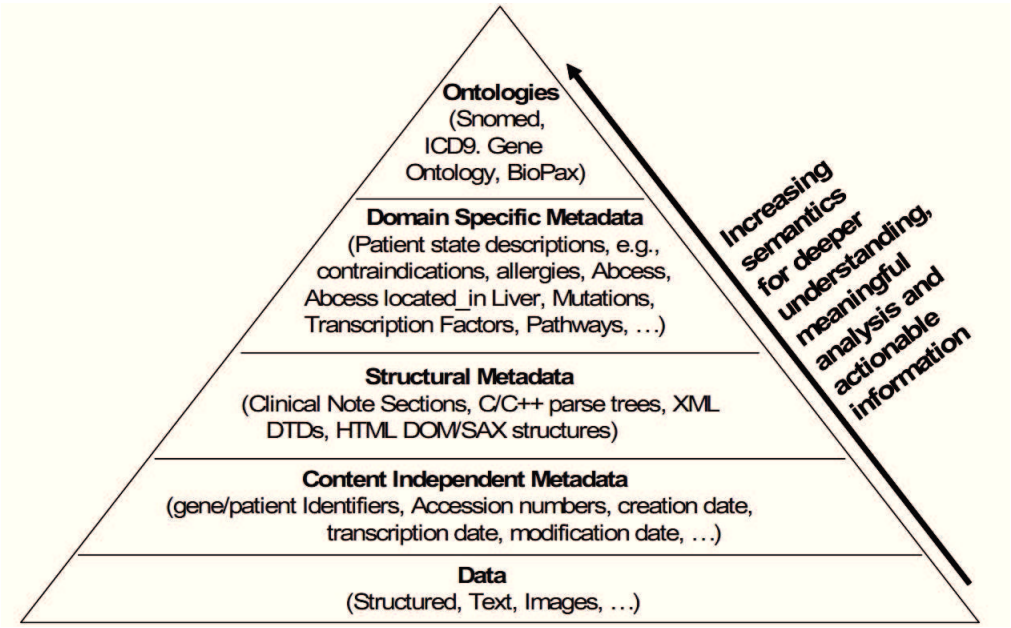
Myšlenka sémantického webu byla představena v roce 2001 a byla prezentována jako rozšíření současného webu, ve kterém se informacím přidělí rovněž jejich význam. Jedná se doplnění stávajícího obsahu o metadata, která daný obsah popisují. Byla zavedena rovnice:

$$Sémantický\ web = data + metadata$$

Definice 2.1 *Metadata jsou definována jako dodatečné informace o datech. Jejich úkolem je lépe popisovat obsah dat. Metadata lze automatizovaně zpracovávat a mají význam při vyhledávání. Mohou popisovat video, obrázky a další multimédia, uplatnění našla například také v geografických systémech či digitálních knihovnách. Rozdělení metadat výborně popisuje obrázek 1 z knihy [31].*

Rozlišujeme metadata:

- *Obsahově nezávislá - nemají přímou závislost k obsahu dat, která popisují. Jedná se například o různé ID, časové známky, a další.*
- *Obsahově závislá - jsou závislá na popisovaném obsahu dat. Zachytávají význam obsahu a jeho vztahy.*



Obrázek 1: Rozdělení metadat

Sémantické informace se neuchovávají pouze v HTML dokumentech, ale využívají se také v rámci ontologií, které se zapisují v jazycích RDF,OWL či XTM.

Definice 2.2 *Ontologií rozumíme množinu konceptů a definování vztahů mezi nimi v dané aplikační doméně. Konceptem je výraz popisující obsah. Často se vyváří celé slovníky konceptů popisující doménu webu. Místo slova ontologie se můžeme setkat také s výrazem* prostor konceptů (concept space, CS).

Dalším pojmem, který se v souvislosti se sémantickým webem hojně využívá je Folksonomie.

Definice 2.3 *Folksonomie je též nazývána sociálním tagováním. Je výsledkem tagování obsahu pro osobní potřebu. Tag můžeme definovat jako vlastní klíčové slovo popisující obsah. Tagování je děláno člověkem a je dostupné i pro ostatní uživatele. Lidé při tagování používají svůj vlastní slovník. Je zdrojem pro chybějící metadata [13].*

Sémantický web je dále popsán v [31].

2.2 Adaptivní web

Adaptivní web vychází ze sémantického webu a rozšiřuje jej o možnost uživatelského přizpůsobení obsahu webu. Sémantické informace při adaptaci a personalizaci hrají klíčovou roli. Adaptivní web ukládá všechna data o uživateli a jeho chování v systému a následně je vyhodnocuje s využitím informací v doméně [1]. Výsledkem je adaptace obsahu a navigace.

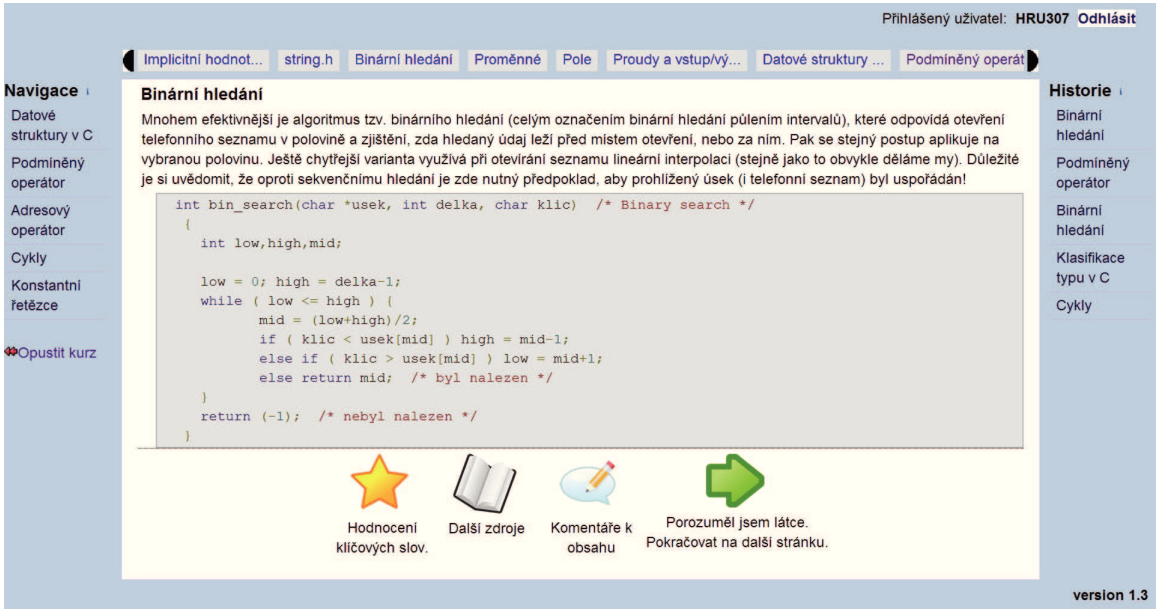
V systému XAPOS je použit model AHAM [9], který definuje tři základní části adaptivního systému:

- *Model domény* popisuje informace v doméně a jejich vzájemné vazby.
- *Model uživatele* obsahuje data o uživateli a jeho chování.
- *Model adaptace* určuje, jak využít modely uživatele a domény k adaptaci obsahu.

Existuje několik adaptivních technik, které výše uvedené modely využívají. Rozdělit je lze na techniky *adaptivní prezentace* a *adaptivní navigace*. Podrobně jsou popsány v [18]. Jak již názvy napovídají, jedná se o adaptaci obsahu či navigace. Častým příkladem adaptivního personalizovaného webu jsou e-learningové systémy. Využití adaptace v této doméně je popsáno v [6].

3 Experiment v systému XAPOS

Pro náš experiment zaměřený na analýzu chování uživatelů adaptivního webu bylo potřeba najít takový systém, který bude možné modifikovat, přivést do něj alespoň několik desítek uživatelů a získat potřebná data k dalšímu zpracování a vyhodnocení. Tyto podmínky splňuje e-learningový systém XAPOS[3] [27], který zhotovil při své disertační práci Ing. Zdenek Velart, Ph.D. XAPOS je používán v doméně VŠB-TU Ostrava a slouží jako experimentální systém. Obrázek 2 slouží pro ilustraci jedné ze stránek systému.



Obrázek 2: Kurz Programování v C/C++ v systému XAPOS

Do systému byly zapracovány změny potřebné pro kvalitní analýzu dat v rámci rozsahu této práce. Jednalo se o sběr komentářů k obsahu, přidávání odkazů na další zdroje (anotací) a také hodnocení anotací s možnosti výběru typu hodnotící stupnice. Po úspěšné implementaci a otestování došlo ke spuštění systému a jeho zpřístupnění studentům prvního ročníku oboru Informatika a výpočetní technika na VŠB-TU Ostrava.

V XAPOSU byl připraven kurz programování v C/C++, který byl připraven pro studenty předmětu *Úvod do programování*. Studenti byli k aktivitě v systému motitováni nabídkou získat bonusové body dle jejich aktivity. Výše zisku bonusových bodů byla sta-

[3] http://arg.vsb.cz/XAPOS/, authors: Zdeněk Velart, Petr Šaloun

novena splněním několika předem připravených podmínek. Potřebnou aktivitu k získání nejméně jednoho bonusového bodu vyvinuly dvě stovky studentů. Díky nim byla získána data potřebná k analýze a ověření hypotéz.

3.1 Systém XAPOS

XAPOS je adaptivní personalizovaný webový systém, který se umí přizpůsobit individualitě svých uživatelů. Systém se zaměřuje na oblast e-learningu a obsahuje také modul pro vytváření testů. Pohyb uživatele v rámci systému a kurzu je pečlivě logován do databáze. Účely systému jsou rovněž experimentální, jelikož díky možnosti rozšiřování je vhodný pro testování hypotéz.

Technicky je systém implementován v jazyku Java s využitím frameworku Struts[4]. Jednotlivé stránky jsou napsány v JSP[5] a využívají také AJAX. Celý systém je zároveň postaven na architektuře MVC. V doméně VŠB-TU Ostrava XAPOS[6] běží na aplikačním serveru Tomcat[7]. Data jsou uložena v dabatabázi MySQL.

XAPOS není možné používat bez přihlášení. Při přihlašování lze použít
kromě systémových údajů také přihlašovací údaje z LDAP. Jelikož se jedná o adaptivní systém implementující architekturu AHAM, tak se XAPOS skládá ze tří základních částí:

- *Model uživatele* - má za úkol uchovávat a spravovat veškeré informace o jednotlivých uživatelích. Jedná se jak o informace, které uživatel sám poskytne, tak o data zjištěna z jeho chování v systému. Evidují se například údaje o průchodu systémem či jednotlivých přihlášeních.

- *Model domény* - je reprezentován obsahem celého kurzu.

- *Model adaptace* - se stará o adaptaci a modifikuje navigaci uživatele. Navigaci se podrobněji věnuje kapitola 3.

XAPOS také s využitím výhod frameworku Struts podporuje vícejazyčnost a to jak systémovou, tak také z hlediska obsahu. Jednotlivé učební texty tak mohou být uloženy v různých jazykových mutacích. Za zmínku stojí fakt, že jednotlivé učební stránky nejsou

[4] http://struts.apache.org/
[5] http://www.jsptut.com/
[6] http://arg.vsb.cz/XAPOS/
[7] http://tomcat.apache.org/

uloženy v databázi ale v souborovém systému a databáze obsahuje jen příslušné reference.

Více se e-learningovým systému a zejména XAPOSu věnuje pramen [24].

3.2 Implementace změn v systému

Pro námi zvolené experimenty či ověření některých hypotéz bylo potřeba do systému implementovat novou funkcionalitu. Po seznámení se systémem, architekturou a jeho zdrojovým kódem bylo navrženo a konzultováno několik změn. Ty byly poté implementovány a otestovány. Jednalo se o:

- Komentáře - prostřednictvím kterých uživatelé vyjadřují svůj názor na obsah a systém. Texty jsou analyzovány pro jejich sentiment.
- Anotace - pomocí kterých uživatelé přidávají odkazy s relevantním obsahem k danému učivu.
- Hodnocení anotací - je důležité pro rozpoznávání kvality zdrojů a je zohledněno při určení pořadí zobrazování anotací k učivu.
- Výběr způsobu hodnocení - dává uživatelům možnost si zvolit stupnici hodnocení, kterou budou pro hodnocení anotací používat.

3.3 Získaná data

Díky dvěma stovkám uživatelů, kteří se v systému aktivně pohybovali, se podařilo získat dostatek relevantních dat. Uživatelé byli k aktivitě motivováni možným ziskem bonusových bodů. Motivaci studentů během experimentu popisuje ve své diplomové práci M. Briš [35]. Tabulka 1 slouží jen pro přehled aktivity dvou stovek studentů. Získaná data jsou podrobněji rozpracována v dalších kapitolách.

3.3.1 Rozlišení kvality získaných dat

Pro potřeby kvalitní analýzy dat a posouzení jejich významu je důležité zajistit relevantnost dat. Jelikož studenti byli k aktivitě motivování získáním bonusových bodů do předmětu hrozilo, že jejich aktivita bude konaná nerelevantně a pouze za účelem zisku bodů. Aby aktivita těchto studentů negativně neovlivnila výsledky analýzy, jsou studenti na základě splnění či nesplnění několika podmínek rozděleni do kategorií.

Aktivita	Celkově
návštěv výukových objektů	38539
hodnocení kl. slov	8466
anotace	1375
hodnocení anotací	3376
komentáře	1471
vykonané testy	2347

Tabulka 1: Aktivita uživatelů v systému

Definovány jsou dvě kategorie studentů - *STUDALL* a *STUDGOOD*. Ve skupině *STUDALL* je všech 200 studentů aktivních v systému XAPOS. Uživatelé, kteří prošli podmínkami na kontrolu vysoké relevantnosti, jsou rovněž také zařazeni v kategorii *STUDGOOD*. Při analyzování dat pak byly zkoumány obě kategorie zvlášť. Toto rozdělení má skutečně velký význam, protože je důležité, aby byly klíčové závěry vyvozené z dat studentů, kteří se systémem pracovali poctivě, tedy z kategorie *STUDGOOD*.

Na začátku se předpokládalo, že všichni studenti patří do skupiny *STUDGOOD*. Definováno bylo devět podmínek pro vyřazení z této skupiny. Při naplnění více než jedné z nich byl uživatel ze skupiny kvalitnějších uživatelů vyřazen. Podmínky vyřazení jsou definovány v tabulce 2.

Pro kontrolu splnění podmínek bylo využito SQL dotazů, které byly implementovány do speciálně připravené aplikace v .NET/C#. V první fázi dochází k načtení všech uživatelů a poté k postupnému ověřování a zapisování trestných bodů za každou splněnou podmínku. Čtyři podmínky se týkají komentářů, protože v nichž se studenti nepatřící do *STUDGOOD* provinili nejvíce. Aby došlo k rozpoznání uživatele, který na prvních stránkách vykoná aktivitu potřebnou k zisku bodů a zbytek stránek již pouze rychle projde, kontrolujeme procentuální četnost ve 20 % po sobě navštívených stránkách. Pokud přesáhne určitou mez, uživatel rovněž obdrží trestný bod. Jak byla kontrola prováděna demonstruje níže uvedená část kódu.

Podmínka vyřazení	Splnilo uživatelů
navštívena méně než 1/2 všech objektů	64
více než 1/3 návštěv stránek s dobou kratší než 10s	9
více než 90 % všech přidaných anotací ve 20 % sekvenčně navštívených objektů	54
více než 90 % všech přidaných hodnoceni anotací ve 20 % sekvenčně navštívených objektů	49
více než 90 % všech přidaných hodnoceni kl. ve 20 % sekvenčně navštívených objektů	39
více než 90 % všech přidaných komentářů ve 20 % sekvenčně navštívených objektů	58
vloženy více než dva úplně stejné komentáře	19
více než pět komentářů kratších 5 znaků	7
žádný komentář delší než 10 znaků	6

Tabulka 2: Podmínky pro vyřazení ze skupiny *STUDGOOD*

```
int lo20 = (Int16) Math.Floor(slr.Count * 0.2); // 20% vzdělávacích objektů
short max_ann = (Int16) Math.Floor(ann * 0.9); // 90% přidaných anotací

  // součet anotací v pvních 20% navštívených objektů
  for (int j = 0; j <= lo20; j++)
  {
    sum_ann += slr[j].ann;
  }

  for (int i = lo20; i < (slr.Count); i++)
  {

    if (i != lo20)
    {
      // odečtení počtu anotací již započítaného objektu
      sum_ann -= slr[i-lo20].ann;
    }

    // přidání počtu anotací dalšího navštíveného objektu
    sum_ann += slr[i].ann;
```

```
        // pokud je počet anotací větší než 90% z celkového počtu
        if (sum_ann > max_ann)
        {
            bad_ann = 1;
        }

    }
```

Výpis 1: Kontrola počtu přidaných anotací ve 20 % sekvenčně jdoucích objektů

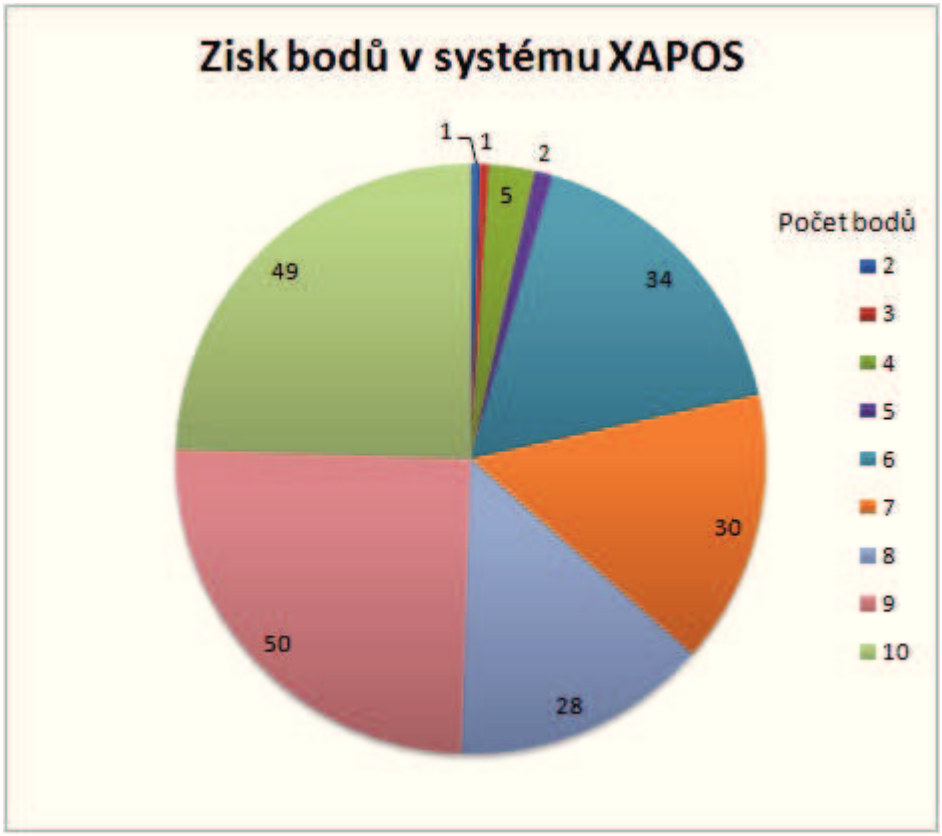
Počet trestných bodů je uložen v databázi a spadá do modelu uživatele. Tolerovatelná hranice, která může být způsobena například nevhodně zvolenou podmínkou pro některé uživatele, je jeden bod. Studenti mající více než jeden trestný bod, tedy splnili alespoň dvě podmínky z tabulky 2, jsou ze skupiny *STUDGOOD* vyřazeni. Celkem bylo vyřazeno 88 studentů. Ve skupině kvalitnějších uživatelů zůstalo 122 uživatelů. Tabulka 3 popisuje počty studentů a trestných bodů podrobněji.

Počet trestných bodů	**Počet uživatelů**
0	40
1	72
2	44
3	34
4	7
5	3
> 5	0

Tabulka 3: Počty trestných bodů při určování skupiny uživatelů *STUDGOOD*

3.3.2 Vztah mezi body získanými v XAPOSu a v předmětu

Z hlediska analýzy chování uživatelů v systému XAPOS je možné stanovit následující hypotézu: *Existuje vztah mezi výsledky studentů v systému a ve studijním předmětu?* Pro tento výzkum byl využit export výsledků studentů ze systému Edison. V XAPOSu bylo možno získat až deset bonusových bodů. Téměř polovina studentů získala 9 či 10 bodů. Naopak 2-5 bodů získalo pouhých 5 % aktivních uživatelů, zbytek dosáhl úrovně 6-8 bodů. Z předmětu je možné získat známku výborně (86-100 bodů), velmi dobře (66-85 bodů) a dobře (51-65 bodů). Studenti, kteří nezískají více než 50 bodů mají známku nedostatečně. Obrázek 3 zobrazuje zisky bonusových bodů.



Obrázek 3: Zisk bodů v systému XAPOS

Při srovnání výsledků v systémech XAPOS a EDISON jsme byli schopni vysledovat určitý vztah. Uživatelé, kteří získali 9 či 10 bodů, získali v předmětu průměrně 68.18

bodu. Naopak ti, co získali maximálně pět bodů za aktivitu, tak v předmětu byli oceněni průměrně 41.67 body. Blíže data popisuje tabulka 4.

Bonusové body	**Počet studentů**	**výborně**	**velmi dobře**	**dobře**	**Průměr bodů**
0-5	99	1	2	2	41.67
6-8	92	11	45	20	59.94
9-10	9	27	45	14	68.18

Tabulka 4: Počty získaných bodů v systému XAPOS a EDISON

4 Navigace v systému

Navigace je jedním z nejdůležitějších prvků adaptivního webu. Lze k ní přistoupit mnoha různými způsoby. Uživatel může mít při průchodu systémem naprostou volnost a využít kompletní navigaci napříč celým webem, může mu být poskytnut přímo výsledek na základě jeho vyhledávacích požadavků nebo lze navigaci založit například na zkušenostech z chování jiných uživatelů. Je důležité zmínit, že navigace je úzce svázána s obsahem a jeho kvalitou. Změna v obsahu může změnit chování uživatelů na webu a může mít následně dopad také na navigaci.

4.1 Adaptace průchodu systému v závislosti na profilu uživatele

Experimentální systém XAPOS, který byl využit také pro tuto práci, má v sobě již zaveden určitý princip navigace. Ten je popsán v [25].

V kontextu XAPOSu je stránka nazývána výukovým objektem a těch je v kurzu *Programování v C/C++* celkem 123. Každý výukový objekt má definovánu svou množinu konceptů, jenž popisují danou stránku. Přičemž jeden koncept může být současně obsažen ve více výukových objektech. Konceptem rozumíme výraz jasně popisující obsah.

V uživatelském modelu je sledován pohyb uživatele v systému a hlavně jsou uchovávány dosažené znalosti. Ty jsou uchovávány prostřednictvím množinou již naučených konceptů nazývanou *množina dosažených znalostí*. Každý koncept se v ní pro daného uživatele vyskytuje maximálně jednou. Jsou tak k dispozici informace o tom, které koncepty se již student naučil.

Navigace uživatele funguje efektivním způsobem. Uživatel se dostane na některou ze stránek kurzu, nastuduje látku a potvrdí její nastudování tlačítkem *Porozuměl jsem látce, pokračovat na další stránku*. Koncepty, které jsou s daným výukovým objektem svázány, jsou uloženy do množiny dosažených znalostí. Pokud uživatel pokračuje pomocí výběru jiného objektu z menu, není množina dosažených znalostí upravena.

Poté dochází k úpravě nastavení adaptace a navigace. Navigační menu se upraví dle množiny dostupných výukových objektů a množiny znalostí daného uživatele. K tomu se využívá některá z metrik popsaných v [18].

4.2 Typy uživatelů

Pro změnu v navigaci je ale vhodné znát, jak se uživatelé na daném webu chovají a jakým způsobem stránky procházejí. Zejména v oblasti e-learningu je důležité znát typ uživatele z hlediska jeho pohybu v systému a přístupu k navigaci. Z analýzy chování uživatelů lze vyvodit změny, které mohou zlepšit navigaci napříč celým systémem, případně daným výukovým kurzem. Existují také algoritmy, které dokážou předpokládat další krok uživatele na webu [26].

V práci [34] byly zkoumány typy uživatelů z hlediska jejich přístupu k navigaci v kvízech. Bylo bráno v úvahu, že v systému je několik kurzů a student se může věnovat i více kurzům najednou. Z pohledu průchodu uživatelů v systému byly definovány následující skupiny chování v systému:

- *Sekvenční průchod* - uživatel prochází kvízy sekvenčně tak, jak jsou mu nabízeny.
- *Opakování* - uživatel po neúspěšném pokusu kvíz ihned opakuje.
- *Vracení se* - uživatel se vrací na předchozí kvíz
- *Přeskakování v kurzu* - uživatel přeskočí na další kvíz v rámci jednoho kurzu
- *Přeskakování mimo kurz* - uživatel přejde ke kvízu v jiném kurzu
- *Skok do pokročilejších kurzů* - uživatel přeskočí ke kvízu v následujících kurzech
- *Skok do pedchozích kurzů* - uživatel přeskočí ke kvízu z dřívějšího kurzu

4.3 Vizualizace průchodu systémem

Při analýze chování uživatelů v prostředí webu má svůj opodstatněný význam rovněž vizualizace. Díky ní je možné názorně vidět, jak se uživatelé v systému pohybují, které stránky odkud navštěvují a na které se nedostanou. Vizualizovány jsou přechody mezi jednotlivými stránkami systému. Ty jsou v případě XAPOSu ukládány do databáze MySQL a mohou být dále zpracovávány vizualizačními nástroji.

Výsledkem vizualizace je orientovaný graf s uzly. Tyto znázorněné uzly odpovídají jednotlivým stránkám (výukovým objektům) a hranami znázorňují jednotlivé přechody mezi stránkami. Hrany mohou být ohodnoceny tak, že jejich hodnota značí počet přechodů mezi danými uzly.

4.3.1 Software pro vizualizaci

Existuje řada knihoven, ve kterých lze provádět vizualizaci přímo ve webové aplikaci. Jednou z nich je například JUNG (Java Universal Network Graf Framework) [8], který je postaven na jazyce JAVA. Podobným projektem je Ravis[9] založený na platformě Flash. Na výběr je také z mnoha dalších vizualizačních nástrojů.

Při hledání vhodné aplikace pro vizualizaci byl kladen důraz zejména na jednoduchost, použitelnost a nezávislost na platformě zdroje dat. To vše při zachování *open source* licence. Vhodným řešením se stal vizualizační nástroj Gephi[10], jehož motto je *makes graphs handy*. Jedná se o moderní vizualizační nástroj s multiplatformní podporou. Samotná aplikace nabízí široké možnosti vizualizace, nastavení a mohou do ní být instalována další rozšíření. Projekt je popsán ve zdroji [36].

4.3.2 Implementace

Aby mohla být vizualizace provedena, je třeba poskytnout aplikaci data ve správném formátu, kterým v tomto případě jsou **.csv* soubory. Pro tento účel byla vytvořena .NET/C# aplikace, jejímž výstupem jsou dva soubory.

- *nodes.csv* - obsahuje seznam uzlů (stránek) ve formátu: *ID; název stránky*
- *edges.csv* - obsahuje hrany (přechody) mezi jednotlivými uzly ve formátu: *zdrojový uzel; cílový uzel;počet přechodů*

Seznam uzlů čítal 123 výukových objektů. Hran bylo v souboru uloženo 3880, kdy číslo vyjadřuje také počet různých využitých přechodů mezi stránkami XAPOSu. Ve vytvořené aplikaci bylo pamatováno také na možnost exportu hran pro jednotlivé uživatele.

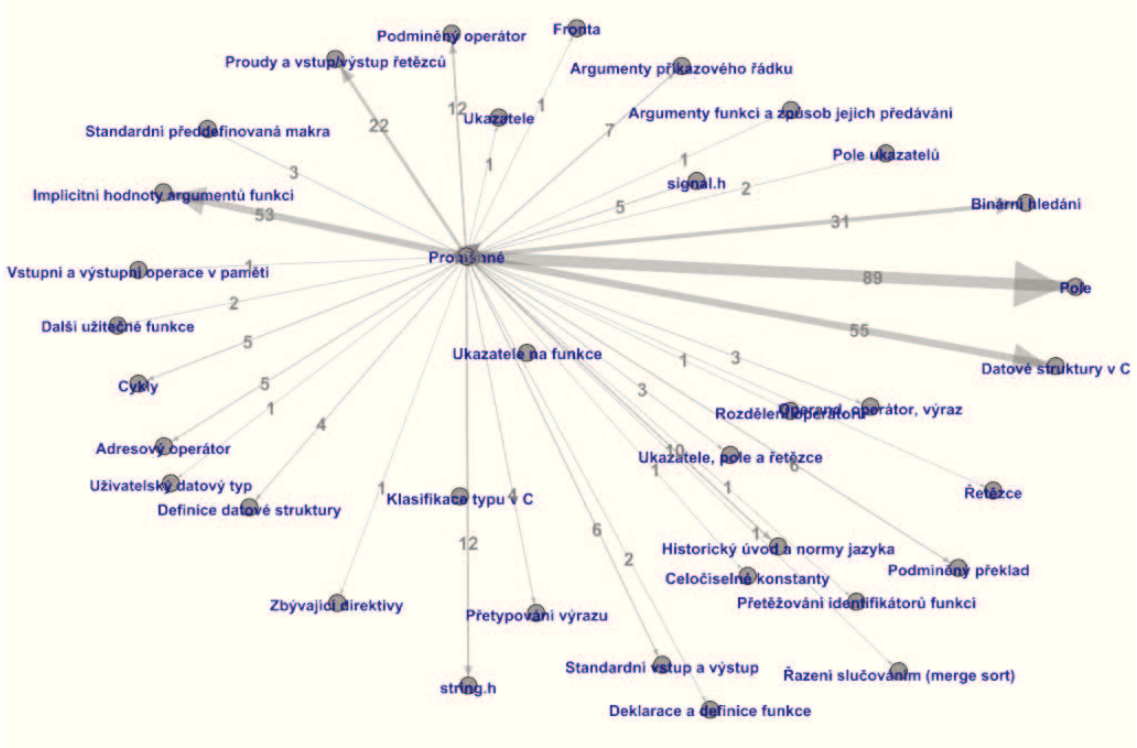
4.3.3 Výsledky vizualizace

Po importu dat do aplikace a nastavení vlastností grafu je možné vygenerovat kompletní vizualizaci průchodů systémem, včetně zesílení hran s větším počtem přechodů. Z hlediska webového systému má smysl vizualizovat cestu webem jednoho uživatele. Z takového grafu lze rozpoznat, jak systémem procházel a jaké stránky a kolikrát navštívil.

[8] http://jung.sourceforge.net
[9] https://code.google.com/p/birdeye/wiki/RaVis
[10] http://gephi.org/

Smysl má také vizualizace pro určitou stránku, kdy je možnost sledovat, z jakých stránek uživatelé na cílovou stránku přicházeli či naopak, kde vedla jejich cesta po shlédnutí obsahu.

Z důvodu velikosti systému a následně také celé vizualizace je v rámci této práce přiložena jen vizualizace pro výukový objekt „Proměnné“. Jsou na ní vidět přechody na další objekty e-learningového systému XAPOS. Vhodnější je ale vizualizace přímo v aplikaci. Obrázek 4 zde slouží pro ilustraci vizualizace části systému. Je na něm zobrazen jeden z výukových objektů a znázorněno, na jaké výukové objekty následně uživatelé přecházeli.



Obrázek 4: Vizualizace pro výukový objekt

4.4 Možnosti ovlivnění navigace

Znalost pohybu uživatelů v systému je pro rozhodování o změnách navigace velice důležitá. Pokud je předpokládán nebo dokonce doporučen určitý pohyb po systému, lze jej srov-

nat s reálným pohybem uživatelů. Jednou z možností je znázornění přechodů mezi stránkami pomocí vizualizace. Díky ní lze lépe pochopit a zmapovat cestu uživatele po systému. Následně je možné chování uživatelů vyhodnotit a učinit případné změny v systému navigace nebo samotném obsahu.

Dalšími údaji vhodnými ke sledování uživatelů v systému jsou strávený čas na jednotlivých stránkách a počet návratů na jednotlivé stránky. Pokud má stránka vysoký počet návratů stejným uživateli a přitom patří mezi běžné stránky systému, můžou vyšší hodnoty ukazovat na problém v navigaci. Nízká doba průměrné návštěvy může naznačovat také možný problém s obsahem.

Při analýze dat v XAPOSu bylo zjištěno, že některé stránky, konkrétně „proměnné", „pole" a „binární hledání", mají vyšší počet opakovaných návštěv stejnými uživateli než ostatní. Při pohledu na strávený čas uživatelů na daných stránkách bylo analyzováno, že studenti tráví na těchto stránkách většinou méně než deset vteřin. Toto zjištění indikuje možný problém v navigaci nebo obsahu učiva.

Navigace XAPOSu je postavena na vyhodnocování konceptů a jejich vazeb vůči množině dosažených znalostí. Správnost modelu konceptů tedy má přímý vliv na navigování uživatele v systému. Na jednotlivé koncepty jsou vázána klíčová slova. Uživatelé v systému komentují obsah a někdy při hodnocení použijí klíčové slovo. Takové komentáře lze analyzovat pro jejich sentiment nebo-li názor. V případě, že jsou klíčová slova určitého konceptu hodnocena v komentářích kladně, pak je koncept použit správně.

Koncepty s negativním sentimentem mohou být potenciálním problémem a podnětem ke zvážení správnosti modelu konceptů a celé navigace. Daný koncept by měl být přezkoumán a případně přehodnocen. Úprava modelu konceptů pro systém XAPOS je posána v [18].

Výsledek analýzy sentimentu komentářů pro určitou stránku systému napovídá, jak jsou uživatelé spokojeni s poskytovaným obsahem. Analyzování emocí z komentářů uživatelů v prostředí webů různých domén může pomoci získat informace využitelné pro kvalitnější doručování obsahu. Díky informacím, co uživateli se uživateli líbí a co ne, lze vylepšit adaptaci a personalizaci.

5 Analýza sentimentu v komentářích

Analýza sentimentu je aktuálně jedno z nejdiskutovanějších témat při zpracovávání informací z webu. V angličtině se můžeme setkat s termíny jako sentiment analysis a opinion mining. Pokud bychom hledali tu nejjednodušší definici sentimentu, stačilo by použít slovo emoce. Pod pojmem analýza sentimentu tedy rozumíme rozpoznávání emocí v textech. Emocí je celá řada, například se může jednat o smutek, strach, štěstí a další. V nejjednodušší formě můžeme sentiment rozdělit na negativní a pozitivní.

V minulosti již bylo vytvořeno mnoho startupů a projektů s jedním jediným cílem - získat sentiment z textu. To, že analýza sentimentu má ve světě internetu skutečně velký význam dokazují také široké možnosti využití. Jedním z největších příkladů je finanční trh, kde analýza sentimentu pomáhá při každodenním obchodování s akciemi. Analyzovány jsou zmínky o produktech či přímo firmách na oficiálních stránkách, blozích, ekonomicky zaměřených serverech a v poslední době také stále více na sociálních sítích, zejména twitteru. Na základě těchto nesmírně cenných informací pak obchodníci s akciemi předpokládají jejich vývoj [3]. V současné době je pro angličtinu funkční řada dostupných řešení, které tyto informace svým klientům poskytují v reálném čase. Jedná se například o SNTMNT[11], OPFINE[12] či Stock Sonar[13].

Analýza sentimentu se rovněž úspěšně uplatňuje v oblasti prodeje produktů a služeb. Zde má velký význam jak pro prodejce, tak také pro zákazníky. Výrobcům či prodejcům může analýza sentimentu pomoci analyzovat zpětnou vazbu na jejich produkty a být podnětem pro zlepšení či různé změny. Výrobci potřebují k dosažení úspěchu pochopit, jak zákazníci na jejich výrobek či službu reagují, co se jim líbí a co naopak lidi od nákupu odrazuje. Potenciální klienti z analýzy hodnocení či recenze znají silné a slabé stránky produktu, což jim napomáhá při rozhodování o koupi. Jako budoucnost optimalizace internetových reklam vidí analýzu sentimentu v článku [33], který je zakončen větou „Neexistuje lepší sentiment než mít peníze zákazníků na svém bankovním účtu."

Metody analýzy sentimentu se dají aplikovat například také na agregátory služeb v oblasti ubytování či e-shopy. Možnosti využití jsou téměř neomezené. V současné chvíli navíc již jsou známy algoritmy, které dokážou detekovat sentiment nejen ze subjektivního textu jasně vyjadřujícího názor, ale také z textu objektivního. Na internetu se každou

[11] www.sntmnt.com/
[12] www.opfine.com/
[13] http://www.thestocksonar.com/

chvíli objeví velké množství článků, komentářů, textových příspěvků či tweetů, které stojí za to analyzovat.

Je důležité zmínit, že analýza sentimentu je NLP (Natural Language Processing) problém a úspěch jednotlivých metod analýzy sentimentu tak závisí na kvalitě dalších NLP metod pro daný jazyk [8, 14].

Existuje mnoho již ověřených metod pro analýzu sentimentu anglicky psaného textu. Velký pokrok byl zaznamenán pro texty zejména ve španělštině a čínštině. Pro většinu dalších jazyků se zatím nejčastěji využívá postup s využitím překladu textu do angličtiny a následné analýzy sentimentu. Vzhledem k tomu, že každý jazyk má řadu svých specifických rysů a ani překlad nemusí být proveden vždy správně, není tento způsob analýzy optimální.

Při návrhu metod a algoritmu pro analýzu sentimentu česky psaného textu jsme se zaměřili na využití různých postupů, které se provádějí rovněž v angličtině, a také na zohlednění rysů speciálních pro češtinu. Hledání emocí v česky psaných textech se v současnosti věnuje jen diplomová práce Radka Července, který však využil metodu strojového učení a nezohledňuje aspekty nalezené v textu [2].

Že analýza emocí v česky psaných textech není nic jednoduchého je patrné z výzkumu Josefa Šlerky, jenž je šéfem vývoje a výzkumu ve společnosti Ataxo Interactive a zároveň také vede Studia nových médií na FF UK. Ten ve své práci dokázal, že při hodnocení sentimentu krátkých příspěvků mají samotní uživatelé problém se shodnout, jaký sentiment vlastně text vyjadřuje. Shoda ohledně třídy sentimentu vyšší než 70 % byla pouze u necelé poloviny příspěvků [7]. Jako závěr pak uvedl: „Lidé mají sami mezi sebou velký problém se shodnout na tom, co je pozitivní, co neutrální a co negativní.“

5.1 Třídy sentimentu

Základní rozdělení sentimentu nebo-li emocí je na pozitivní a negativní. Pokud se v textu sentiment nevyskytuje nebo je ve stejné míře zastoupen pozitivní i negativní sentiment zároveň, můžeme mluvit o sentimentu neutrálním. Při vyhodnocování celkového sentimentu určitého textu bereme v potaz také míru pozitivního či negativního zabarvení. Jako výsledek analýzy sentimentu je zařazení do jedné z emočních tříd. Rozeznáváme negativní, silně negativní, neutrální, pozitivní a silně pozitivní emoce. Speciálním případem je vulgární negativní sentiment. Pokud jsou vulgarismy použity s pozitivními emocemi,

mají za následek zesílení pozitivního sentimentu. V jiném případě jde o sentiment vulgární. Emoční třídy jsou i s příklady definovány v tabulce 5

emoční třída	**popis**	**příklad**
vulgar	Vulgární neg. emoce	„Koupě toho zasraného foťáku byla chyba.“
negative2	Silně negativní emoce	„Nenávidím tento fotoaparát.“
negative	Negativní emoce	„Tenhle fotoaparát není dobrý.“
neutral	Neutrální emoce	„Tenhle dobrý fotoaparát má pár malých chyb.“
positive	Pozitivní emoce	„Tenhle foťák můžu doporučit.“
positive2	Silně pozitivní emoce	„Zbožňuji tenhle fotoaparát.“

Tabulka 5: Emoční třídy

5.2 Způsoby analýzy sentimentu

Sentiment může být určován na třech úrovních [12]:

Rozsah dokumentu (Document Level): Výsledkem této analýzy je určení pozitivního, neutrálního či negativního sentimentu pro celý dokument. Přitom ale předpokládáme, že celý dokument je vztažen k pouze jedné entitě (aspektu) pro kterou hledáme sentiment.

Rozsah věty (Sentence level): Na této úrovní je určováno, zda jednotlivá věta vyjadřuje pozitivní, negativní či neutrální názor.

Rozsah aspektů (Entity and Aspect level): Předchozí dvě úrovně neurčují přesně, co se autorovi líbí či nelíbí, ale jen sdělují líbí/nelíbí. Analýza sentimentu jednotlivých aspektů je postavena na pravidlu, že každý názor se skládá ze sentimentu (pozitivní/negativní) a objektu (hodnocený objekt). Tato úroveň určení emocí je ze všech nejpodrobnější a nejsložitější, ale zároveň přináší ty nejcennější informace. Jsme totiž schopni určit, kterého aspektu se vyjádřený sentiment přesně týká.

K analýze sentimentu se využívá několik odlišných postupů, a tak můžeme algoritmy rozdělit do skupin na: slovníkový přístup, strojové učení a hybridní metody.

5.2.1 Slovníkový přístup

Nejčastěji se algoritmy pro analýzu sentimentu opírají o slovníkový přístup. Jak z názvu vyplývá, základem metody je slovník frází, na němž je úspěšnost celé metody přímo

závislá. Slovník nemusí obsahovat pouze jednotlivá slova, ale zároveň celé fráze, které zpřesňují klasifikaci emoce. Fráze jsou poté vyhledávány ve slovnících.

Slovník se vytváří třemi možnými způsoby:

- manuálním vkládáním.
- pomocí jiného slovníku (např.: WordNet).

Jedna z metod je založena na slovníku WordNet[14], který obsahuje vztahy mezi jednotlivými slovy. Funguje na principu zjišťování nejkratší vzdálenosti vztahů mezi slovy. Jako základní slova byla určena dobrý a špatný. Při zjišťování sentimentu například pro slovo štědrý je vypočítaná jeho nejbližší vzdálenost ke slovům dobrý (vzdálenost=2) a špatný (vzdálenost=6). Slovu štědrý je tak definován pozitivní sentiment [4].

5.2.2 Strojové učení

Tato metoda je založena na tréninkové množině a využívá některý z klasifikačních algoritmů jako jsou SVM, Naive Bayes nebo Logistic Regression. Tréninková množina obsahuje již sentimentem ohodnocené věty pro každou z definovaných emočních tříd. Pomocí této metody se ale nejčastěji určují pouze dvě emoční třídy - pozitivní a negativní. Kvalita tréninkové množiny a příbuznost domény obsahu jsou hlavním kritériem pro úspěch metody strojového učení.

5.2.3 Hybridní metody

Spojením výhod slovníkové metody a strojového učení jsou hybridní metody. Ty v prvotní fázi využívají slovník a předzpracování textu, a teprve poté aplikují metodu strojového učení [5].

5.3 Sběr dat určených k analýze sentimentu

Navržený postup analýzy sentimentu je otestován ve dvou naprosto odlišných doménách. První z nich jsou komentáře uživatelů e-learningového systému XAPOS běžícího v doméně vsb.cz, který je pro náš výzkum k dispozici. Od dvou set aktivních uživatelů bylo získáno téměř 1500 komentářů, které byly podrobeny analýze sentimentu. Studenti byli

[14] http://globalwordnet.org/

vyzvání ke komentování jak vlastností celého systému, tak také zhodnocení samotného obsahu. Pro podporu různorodosti komentářů byly komentáře jiných uživatelů záměrně skryty. Jak studenti přidávali komentáře je ilustrováno na obázku 5.

Obrázek 5: Přidání komentáře v systému XAPOS

Dalším zdrojem dat pro testování úspěšnosti našeho algoritmu se stal server heureka.cz, kde uživatelé hodnotí zboží zakoupené v různých e-shopech na českém internetu. Tato data se ukázala být pro testování a sestavování postupu ideální. Příklad uživatelské recenze na jeden z fotoaparátů na serveru heureka.cz můžete vidět níže na obrázku 6.

5.4 Návrh vlastního řešení analýzy sentimentu

Při návrhu vlastního postupu automatické analýzy sentimentu jsme vycházeli z dobré znalosti češtiny a rozhodli se při návrhu využít obohacený slovníkový přístup s velkým důrazem na aspekty, tedy objekty názoru či emocí. Protože český jazyk je poměrně rozmanitý, byly brány v potaz různé výjimky, které mohou výsledný sentiment výrazně ovlivnit. Pro implementaci a testování vlastního postupu byla zvolena technologie .NET a jazyku C# a jako úložiště dat a slovníku byl použit MS SQL Server 2008.

Hodnocení jednotlivých příspěvků, pro které sentiment analyzujeme, probíhá naprosto izolovaně a nedochází k jakémukoliv vzájemnému ovlivnění. Celý proces analýzy sentimentu lze rozdělit do několika kroků. Nejprve je text předzpracován tak, že se text převede na malá písmena a rozdělí na menší části do vět. Následně se provádí definice *n-gramů*, jenž umožňují výrazně zlepšit úroveň analýzy sentimentu, protože nedochází pouze k porovnávání slov, ale celých frází.

Obrázek 6: Uživatelská recenze na serveru Heureka k fotoaparátu Nikon

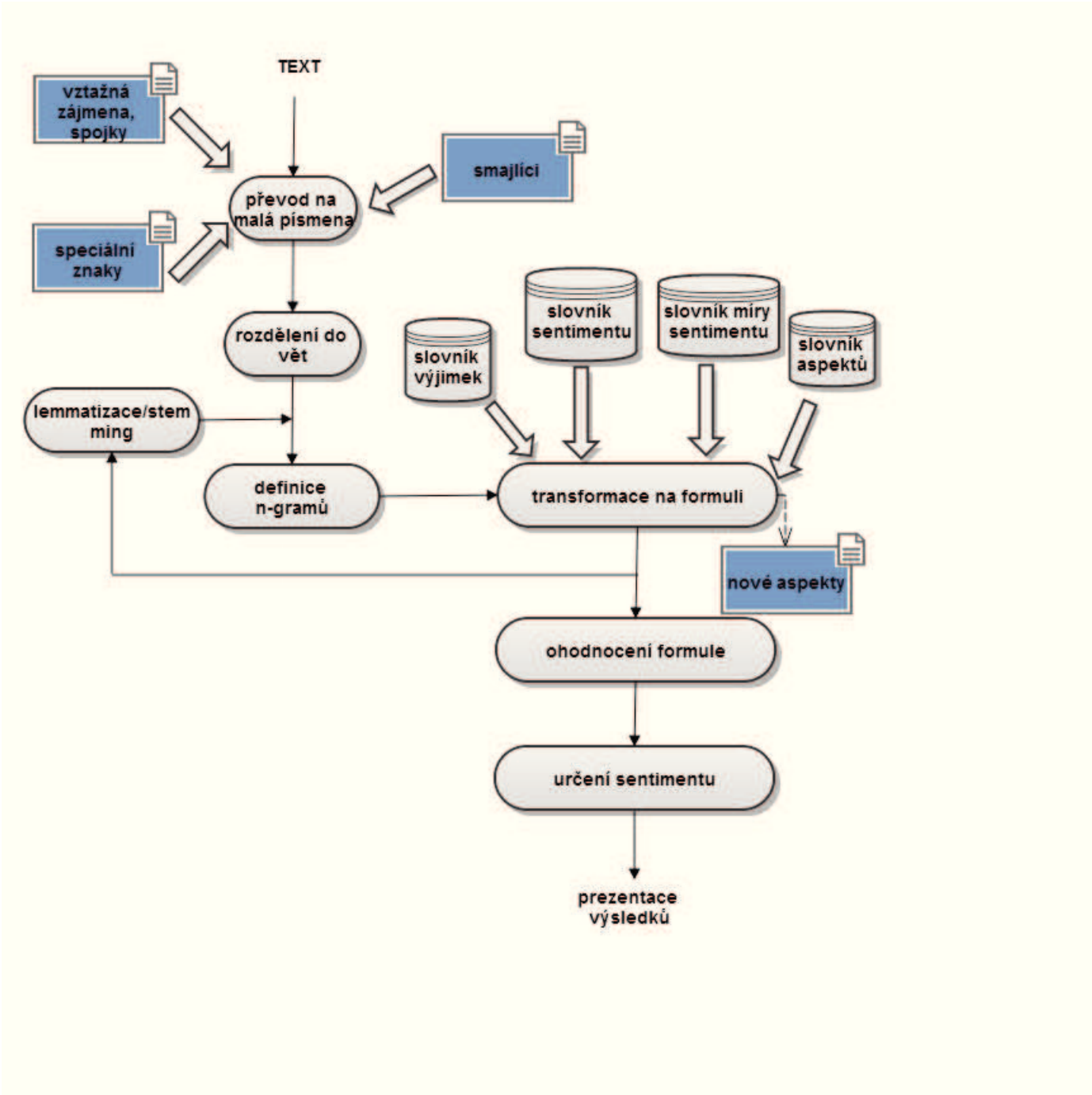
Hlavní částí celého procesu analýzy sentimentu je transformace textu na ohodnocenou formuli. Jednotlivým n-gramům je přiřazen správný význam tak, že dochází k rozlišení a nalezení významu frází vyjadřujících sentiment, míru sentimentu, aspektů a speciálních výrazů. Je to první kolo hledání ve slovnících sentimentu. Důvodem pro provedení prohledávání před upravením slov na základní formu je možnost různého ohodnocení specifických forem slov a také nebezpečí ztráty informace nesprávným převedením tvaru slova.

Dalším krokem je *stemming*, což znamená převedení slov do jejich základního tvaru (stemma). Ten se provádí na slova, která nebyla při prvním hledání rozpoznána. Následuje druhé kolo definice *n-gramů* a transformace textu na formuli. Tentokrát se však již porovnávají *n-gramy* převedené do základního tvaru.

Formule je poté ohodnocena dle předem definovaných pravidel. Z upravené formule lze určit výsledný sentiment a také emoční třídu jednotlivých aspektů.

V posledním kroku se provádí prezentace výsledků a jejich zápis do databáze. Ta může být navržena tak, aby byly uloženy jednotlivé vazby mezi komentářem, analýzou sentimentu a aspekty.

Zjednodušené blokové schéma navržené analýzy sentimentu je na obrázku 7. Jednotlivé části jsou pak podrobně popsány v následujících kapitolách níže.



Obrázek 7: Zjednodušené blokové schéma analýzy sentimentu

5.4.1 Předzpracování textu

Vstupem do analýzy sentimentu je text libovolného rozsahu. Nejčastěji se jedná o text subjektivního typu, ať už se jde o komentář, slovní hodnocení či třeba recenzi.

Pokud je analýza sentimentu součástí systému, tak je vhodné zkontrolovat, zda v databázi již není uložen či dokonce již nebyl analyzován stejný text. Pokud ano, můžeme být příspěvku ihned přiřazen stejný sentiment a případně může dojít k označení potenciálního spamu.

Dalším krokem je převedení textu na malá písmena, to významně usnadní další procesování. Avšak někdy i slovo psané velkými písmeny může mít význam pro analýzu sentimentu. Pokud je tedy text psán malými písmeny a vyskytují se v něm jen některá slova psaná velkými písmeny, označíme toto slovo značkou „{CL}“ a v případě, že se bude jednat o slovo mající vliv na sentiment, bude jeho význam zvýšen.

Zároveň se provádí detekce diakritiky z důvodu zpřesnění správnosti výsledků. Bylo by možné z textu diakritiku odstranit a poté provádět porovnání frází, ale při testování jsme došli k závěru, že lze dosáhnout lepších výsledků, pokud příspěvky obsahující diakritiku budou porovnávány se slovníkovými výrazy obsahujícími diakritická znaménka a naopak. Je však důležité znát, zda se v příspěvku diakritika používá či nikoliv. Toho je dosaženo stanovením minimálního očekávaného podílu znaků s diakritickými znaménky na 5 % všech znaků textu. Aby se předešlo nerozeznáním fráze vinou špatné diakritiky, tak pokud slovo obsahující diakritiku nebude ve slovnících nalezeno, provede se hledání s odstraněním diakritiky.

Provedena je také základní úprava textu, například přidání mezer za interpunkční znaménka. Následnou fází je transformace celého souvislého textu na jednotlivé úseky textu, označme je jako věty. Základem pro toto rozdělení jsou některá interpunkční znaménka a také určité typy spojek. Vzniklé věty pak budou v dalších krocích zpracovávány samostatně. Blíže rozdělující znaky a výrazy popisuje tabulka 6, přičemž jsou seřazeny dle priority pro určování. Například spojka „který“ částečně přenáší aspekt do další věty.

Ke sdílení aspektu z předešlé věty dochází v případě, kdy následná věta začíná ukazovacími zájmeny: „ta, ten, to, toto, tyto“. Sdílení aspektu s jinou větou ale může být nepřesné, a tak je takovému aspektu zvýšena vzdálenost ke slovu vyjadřujícímu sentiment.

znaky/výrazy	**poznámka**
emotikony	nerozdělují věty, jsou přepsány na emoční třídu
';' a '.'	
'!'	zdůraznění sentimentu
'?'	v případě výskytu sentimentu značí emoční třídu negative
kdo, co, jaký, který, čí, jenž	vztažná zájmena rozdělující věty, ale zároveň dochází k částečnému přenesení aspektu do věty následující
a, ale, avšak, však, leč, nýbrž, naopak, jenomže, jenže	spojky rozdělující věty, ale zároveň dochází k částečnému přenesení aspektu do věty následující
','	částečné přenesení aspektu do věty následující

Tabulka 6: Rozdělující znaky a výrazy

5.4.2 Stemming a lemmatizace slov

Čeština patří mezi flektivní jazyky založené na *flekci*, tedy skloňování a časování slov. Při zpracovávní textu tak musíme brát do úvahy tvary slov. K převedení slova na jeho základní tvar se používá stemming či lemmatizace.

Definice 5.1 *Stemming je proces převedení slov do jejich základní formy.*

Definice 5.2 *Lemmatizace přiřazuje slovům jejich základní tvar (lemma) na základě rozpoznání slovního druhu.*

Určení základního tvaru slova je specifické pro každý konkrétní jazyk, je ovšem známo několik obecně platných algoritmů, jak lemmatizaci či stemming provádět. Existují například Brute Force metody, kdy se tvary převádí pomocí rozsáhlé tabulky. Možnosti lemmatizace v česky psaných textech jsou popsány ve zdroji [15].

Rychlým a poměrně účinným algoritmemn který je vhodný pro nalezení základního tvaru slova vyjadřujícího sentiment, je Suffix Stripping algoritmus. Ten je založen na principu přepisování koncovek. Velkou výhodou tohoto přístupu je jednoduchost, avšak může dojít k nerozpoznání nepravidelných slov. Slova vyjadřující sentiment jsou ale většinou pravidelná, takže je Suffix Stripping algoritmus vhodnou volbou [17].

Protože pro účely analýzy sentimentu není slovní druh slova příliš podstatný, je metoda stemmingu naprosto dostačující. Jak provádět stemming v česky psaných tex-

tech je popsáno v pramenu [16]. Je však také důležité brát v potaz jednotlivé dialekty. Nejvýznamější je český pražský a moravský. Z důvodu častého použití byl algoritmus doplněn o další možné koncovky specifické pro jednotlivé dialekty.

Příklad základního tvaru a od něj odvozených rozšířených tvarů pro přídavné jméno a sloveso je v tabulce 7. Základní tvar je poté použit pro nalezení sentimentu.

základní tvar	**odvozené tvary**
hezký	hezké, hezká, hezkou, hezkého, hezkém, hezkému, hezkým, hezkými, hezkých, hezkej, hezku
doporučit	doporuč, doporučuji, doporučil, doporučuje, doporučoval, doporučováno, doporučeno

Tabulka 7: Základní a odvozené tvary slov

5.4.3 Nalezení n-gramů

Čeština obdobně jako další jazyky obsahuje fráze složené z několika slov, které mohou vyjadřovat určité negativní či pozitivní emoce. Při rozložení této fráze na jednotlivá slova a jejch podrobení samostatné analýze sentimentu ale může dojít k určení neutrálního nebo nesprávného sentimentu. Z tohoto důvodu má skutečný význam analyzovat celé slovní obraty a ne jen samotná slova [19]. Při následném hledání významu jednotlivých n-gramů se nejdříve ohodnocují trigramy, poté bigramy a unigramy. N-gramy se využívají zejména při analýze sentimentu pomocí metody strojového učení, ale výhody jejich užití lze čerpat také při slovníkové metodě [20].

Definice 5.3 *n-gram je souvislá posloupnost n slov z textu nebo řeči*
Rozdělení n-gramů:

- *Unigramy: jednotlivá slova.*
- *Bigramy: kombinace všech dvou po sobě jdoucích slov.*
- *Trigramy: kombinace všech tří po sobě jdoucích slov.*

Příklad 5.1
Mějme větu: „Bylo by dobré přidat podporu karet.“

Pokud bychom sentiment analyzovali bez použití n-gramů, tak jediné slovo, které má emoční význam je slovo „dobré“. Výsledkem by byl lehce pozitivní sentiment. Jesliže ale větu rozložíme na n-gramy:

- Unigramy: Bylo, by, dobré, přidat, podporu, karet.
- Bigramy: Bylo by, by dobré, dobré přidat, přidat podporu, podporu karet.
- Trigramy: Bylo by dobré, by dobré přidat, dobré přidat podporu, přidat podporu karet.

Bigram „Bylo by“ je pro nás významný, protože jej při pozdějším zpracovávání nalezneme ve slovníku výjimek a dojde ke změně sentimentu celé věty na lehce negativní sentiment. ■

5.4.4 Transformace na formuli

Před vyhodnocením samotného sentimentu se provádí transformace věty na formuli za pomocí ohodnocení jednotlivých n-gramů. Význam se n-gramům přiřazuje po jeho nalezení ve slovnících sentimentu, míry sentimentu, emotikon, aspektů a výjimek. Výstupem je poté formule, ze které se provádí vyhodnocení a klasifikace sentimentu. Značky a jejich význam jsou v tabulce 8. Složené závorky slouží k jednoznačnému odlišení od ještě neurčených slov ve zpracovávané větě.

Určení případné negace celé věty je velice důležité, protože to má za následek změnu ohodnocení multiplikátoru sentimentu. Důvod, proč vyhledávání výjimek a slov určujících negaci věty probíhá jako první v pořadí demonstruje příklad 5.2

Příklad 5.2
Uvažmuje například slovo „zcela“.
Je rozdíl, zda bude toto slovo použito ve větě „Je to zcela špatně.“ a „Není to zcela špatně.“. V prvním případě je sentiment silně negativní, ve druhém lehce pozitivní. Přičemž M je ohodnoceno 1.5 respektive 0.5 v druhé případě pozitivní věty. Podmínka, která toto určuje, je definována ve slovníku výjimek. ■

Pokud se pro nějaké slovo nepodaří najít trigram, bigram ani unigram v žádném ze slovníků a přitom unigram začíná předponou „ne“ nebo „nej“, tak je potřeba předponu odstranit a pokusit se unigram vyhledat znovu. Pokud je sentiment slova bez předpony

výraz	**význam**
{A#ID}	aspekt a jeho id
{PA#ID}	nový navrhovaný aspekt a jeho id
{CL}	zesílení významu
{!V}	negace sentimentu celé věty
{S}	sloveso určující blízký aspekt
{!}	negace následujícího slova
{Mn}	multiplikátor sentimentu, kde n je n desetinné číslo <0,2>
{VUL}	sentiment třídy vulgar
{NEG2}	sentiment třídy negative2
{NEG}	sentiment třídy negative
{NEU}	sentiment třídy neutral
{POS}	sentiment třídy positive
{POS2}	sentiment třídy positive2

Tabulka 8: Značky transformace

nalezen, tak je pak dle předpony ovlivněn při „ne“ negací sentimentu a v případě předpony „nej“ zdvojnásobením sentimentu daného slova. Pokud slovo s odstraněnou předponou „ne“ není nalezeno, je přidán mírný negativní sentiment celé větě. Tyto změny se projeví do výsledné formule.

Posloupnost kroků je znázorněna na příkladu 5.3.

Příklad 5.3
Zvolme větu: „Fotoaparát není příliš dobrý na focení v tmavých místnostech.“
Postup přepisu na formuli:

- Žádný speciální znak nebyl nalezen.
- Ani jeden z bigramů či trigramů není nalezen ve slovnících, pokračuje se ohodnocováním unigramů.
- Ve slovníku výjimek je nalezeno sloveso není, které je přepsáno na: {S}{!V}.
- Žádný existující aspekt není nalezen.

- Ve slovníku míry sentimentu je nalezeno slovo příliš, které je v případě předchozího nalezení {!V} přepsáno na: {M0.5}.
- Ve slovníku výrazů sentimentu je nalezeno slovo dobrý, které je analyzováno jako positivní sentiment třídy positive a přepsáno na {POS}.
- Jako potenciální aspekt je určeno slovo fotoaparát {PA*ID*}, kde *ID* je buď identifikační číslo z tabulky existujících potenciálních aspektů nebo zcela nové. ID.
- Slova, která se nepodařilo určit, jsou přepsána na $\{n\}$, kde n je počet neurčených slov jdoucích za sebou. V tomto našem případě jich je pět.
- Výsledná formule vypadá takto: „{PA123} {S}{!V} {M0.5} {POS} {5}“.

■

5.4.4.1 Slovník sentimentu

Při analýze sentimentu založené na slovníkové metodě jsou slovníky frází a výrazů nejdůležitější částí metody. Jejich kvalita a rozmanitost má na správnost výsledku největší vliv. Jelikož je čeština velice rozmanitá, nebylo by příliš efektivní vytvářet celý slovník manuálně, i když existují postupy, které jsou na ručním vytváření slovníku sentimentu založeny.

Další možnost vytváření slovníku je založena na strojovém učení, kdy jsou do slovníku postupně ukládána všechna použitá slova a jejich bigramy či trigramy. Při tomto postupu je vhodné na začátku zpracovat a ohodnotit co nejvíce textu. Řada algoritmů nabízí také průběžné ruční ohodnocování nejčastěji používaných ještě neohodnocených n-gramů.

Základem pro třetí užívaný způsob tvorby slovníku jsou synonyma k již uloženým slovům. Na začátku se pro slovník definuje několik slov pro každou třídu sentimentu a poté se provádí hledání synonym. Synonyma, která ještě ve slovníku nejsou, jsou uložena a slouží jako zdroj pro hledání dalších. Často se využívá například databáze slov a vazeb WordNet [21].

Protože nestačí jen znát, zda dané slovo vyjadřuje pozitivní či negativní sentiment, je dobré mít nad slovníkem kontrolu a jeho vytváření poloautomatizovat. K tomu byla vytvořena jednoduchá aplikace v .NET/C#, která má za úkol stahovat ze stránky synonyma-online.cz [15] synonyma způsobem popsaným výše. Web synonyma-online.cz je založen na

[15] http://www.synonyma-online.cz/

slovníku Tezaurus. Pro ukázku v tabulce 9 můžete vidět synonyma nalezená pro slova skvělý, dokonalý a nevhodný.

zdrojové slovo	**synonyma**
skvělý	vynikající, senzační, jedinečný, dokonalý, obdivuhodný, skvostný, úžasný, báječný, nádherný, žárlivý, rozkošný, ohromný
dokonalý	bezvadný, vybraný, vynikající, úplný, perfektní, stoprocentní, ideální, vzorný, naprostý, bezúhonný, skvělý
nevhodný	nešikovný, netaktní, nevyhovující, nepatřičný, neslušný, nemístný, nesprávný, nepřípadný, nehodící se, nezpůsobný

Tabulka 9: Synonyma pro slova vyjadřující emoce

Každému novému nalezenému slovu je automaticky přiřazena stejná emoční třída, jaká je určena ke zdrojovému slovu. Pro správnost slovníku je ovšem vhodné správnost určené emoční třídy zkontrolovat manuálně. Databáze synonym totiž obsahuje vazby napříč třídami positive a positive2 či negative a negative2.

5.4.4.2 Slovník míry sentimentu

Pro analýzu sentimentu mají velký význam také n-gramy vyjadřující míru sentimentu. Patří mezi ně například: zcela, naprosto, částečně. I pro tato slova je možné použít slovník synonym a obohatit tak slovník míry sentimentu o řadu nových výrazů. Díky velké podobnosti ale tato slova vyžadují ruční zadání desetinné hodnoty násobku sentimentu, která se pohybuje na uzavřeném intervalu <0,2>. Přičemž ve slovníku uchováváme multiplikační hodnotu po použití ve větě s pozitivním i negativním slovesem. Důvod používání různých multiplikací byl demonstrován v příkladu 5.2.

V tabulce 10 jsou pro ilustraci uvedena synonyma nalezená pro slova zcela a trochu.

zdrojové slovo	**synonyma**
zcela	jedině, úplně, výhradně, plně, čistě, naprosto, absolutně, docela
trocha	špetka, troška

Tabulka 10: Synonyma slov vyjadřujících míru sentimentu

5.4.4.3 Slovník výjimek

Slovníky slov sentimentu a míry sentimentu jsou účinné, ovšem čeština je natolik rozmanitý jazyk, že v mnohých situacích pro správné určení sentimentu tyto slovníky nestačí. Z tohoto důvodu musí být analýza sentimentu rozšířena o slovník výjimek. Tento slovník obsahuje n-gramy, které mohou mít jakýkoliv z významů definovaných v tabulce 8. Často se v něm vyskytují například n-gramy slovesných spojení, které mají zásadní vliv na výsledný větný sentiment.

Je vhodné doplnit například také slangové výrazy a některé často používané cizojazyčné fráze. Například zkratka „gj“ znamená „good job“ nebo-li „dobrá práce“ a je hojně využívána v diskusích na českém internetu. Sentiment této zkratky je důležité vzít v potaz a slovník výjimek je správným místem pro její uložení. Stejně tak slovník obsahuje také vulgarismy, které mají vliv na konečný sentiment.

Ve slovníku výjimek ale mohou být definovány také různé podmínky, za kterých se slovu přiřadí do formule daný význam. Za podmínku lze považovat předešlé prvky formule či výskyt symbolu {!V}, znamenajícího negaci celé věty, ve formuli. Příklad číslo 5.4 ukazuje, jak může být použití podmínek významově důležité. Tabulka 11 obsahuje některá slova ze slovníku a jejich význam.

Příklad 5.4
Slovo „boží“ má dva odlišné významy dle použití v konkrétní větě. Můžeme jej užít ve větě „Nedal na slovo boží.“, ale také ve větě „Tento článek je boží.“. Ve druhém případě slovo vyjadřuje emoce, v prvním případě však nikoliv.
Pro ohodnocení slova „boží“ emoční třídou positive2 tedy stanovíme podmínku, že tomuto slovu musí předcházet sloveso „{S}“ nebo multiplikátor „{M#}“. V ostatních případech nebude mít dané slovo na sentiment žádný význam. Tato podmínka bude uložena ve slovníku a dle jejího splnění se případně vybere daný význam. ■

n-gram	význam
bylo by vhodné	{S}{M0.5}{NEG}
není	{S}{!}
cool	{/POS}

Tabulka 11: Význam některých n-gramů ve slovníku výjimek

5.4.4.4 Slovník aspektů

Analýza sentimentu zahrnující rozpoznávání aspektů je tou nejkomplikovanější formou dolování emocí z textu. Výsledkem totiž není pouhá emoční třída zkoumané věty, ale dvojice objekt hodnocení a vyjádřený sentiment. Objekt hodnocení nazýváme aspekt. Aspekty jsou uloženy v samostatném slovníku a obsahují rovněž ručně definovaná synonyma jak je ukázáno na příkladu 5.5. Manuálně definované aspekty jsou specifické pro každou doménu, ve které je analýza sentimentu použita [23].

Příklad 5.5

Za vhodný aspekt pro e-learningový systém XAPOS bylo zvoleno slovo „příklaď". Zároveň ale můžeme pro vyjádření stejného významu použít slovo „ukázka" nebo různé skloňování slova „demonstrace", takže jej vložíme do slovníku jako synonyma ke slovu „příklaď". Výsledky sentimentu pro aspekt „příklaď" pak obsahují i data získaná pro synonyma aspektu. Pro názornost lze uvést, že slovo „příklaď" bylo v komentářích použito 133x, „ukázka" 40x a různé verze slova „demonstrace" 4x. ■

Aspekty je velice vhodné využít například u zboží, kde v recenzích a komentářích bývá každá z vlastností produktu či služby hodnocena zvlášť. Na obrázku 8 je uvedeno, jaké základní aspekty byly použity při analýze recenzí fotoaparátu na heureka.cz.

Obrázek 8: Aspekty navržené pro fotopaparát

Aspekty ale nemusí být definovány vždy pro celou doménu, ale také jen pro jednotlivé stránky, kde se analyzované příspěvky vyskytují. Toho je možno využít v systému

XAPOS, kde kromě aspektů určených pro celý systém můžeme definovat aspekty obsaženy v jednotlivých probíraných látkách. V disertační práci Zdeňka Velarta [18] jsou pro každou stránku kurzu určena klíčová slova. Takto definována klíčová slova má smysl využít jako aspekty při analýze sentimentu. Analýza sentimentu poté poskytuje zpětnou vazbu k pojmům obsaženým v učebních textech systému. Ontologií definovaná klíčová slova pro jednu ze stránek a jejich výskyt uvádí tabulka 12.

název objektu	**aspekt**	**pčet výskytů**
Podmíněný operátor	operátor	10
Ukazatele	ukazatel	7

Tabulka 12: Aspekty specifické pro stránku a jejich četnost výskytu v komentářích

Implicitní vazba k aspektu je možností, jak může být prostřednictvím jednoho n-gramu vyjádřen sentiment a také aspekt zároveň. V textu není aspekt zmíněn explicitně ale pouze nepřímo. Při definování takového *n-gramu* je třeba jej uložit do databáze výjimek. Využití implicitní vazby aspektu a slova vyjadřujícího sentiment je ukázáno na příkladu 5.6.

Příklad 5.6
Mějme větu „Fotoaparát je příliš drahý“. I když máme pro doménu definován aspekt „cena“, tak ve větě nebude nalezen. Pokud by větu analyzoval člověk, pak výsledkem bude emoční třída negative pro aspekt „cena“. Jenže v případě automatické analýzy je výsledkem jen emoční třída negative bez vazby na daný aspekt.

Pokud je ovšem do slovníku výjimek přidáno slovo „drahý“, které je ve formuli nahrazeno významem „{NEG} cena“, pak v dalším kroku přepisu na formuli dojde k nalezení aspektu „cena“. Díky existenci impicitní vazby je tedy dosaženo stejného výsledku, jako v případě vyhodnocení člověkem. ■

5.4.5 Vyhodnocení sentimentu

Je důležité rozlišovat jak sentiment vyhodnocený pro jednotlivý aspekt, tak i pro celou větu. V obou případech ale vyhodnocení sentimentu probíhá z formule získané v předešlém kroku algoritmu. Výsledek sentimentu pro celou větu je jedna z emočních tříd definovaných v tabulce 5 kapitoly 5.1 a desetinné číslo z intervalu <-8,8>vyjadřující *emoční koeficient*. Podle něj se pak stanovuje příslušná emoční třída sentimentu. Tabulka

13 znázorňuje emoční třídy, rozsah jejich koeficientu a emoční koeficient slova v dané třídě.

emoční třída	**rozsah výsledného emočního koeficientu**	**emoční koeficient slova**
vulgar	nevyužívá emoční koeficient	nevyužívá emoční koeficient
negative2	<-8,-2>	-2
negative	$(-1.5, -0.25)$	-1
neutral	<-0.25,0.25>	0
positive	$(0.25, 1.5)$	1
positive2	<1.5,8>	2

Tabulka 13: Emoční třídy a jejich koeficienty

Pro určení koeficientu z formule se používá následující postup:

1. Pokud formule obsahuje vulgarismy {VUL} a přitom neobsahuje žádný pozitivní sentiment třídy {POS} či {POS2}, pak je větě určena emoční třída vulgar a analýza věty končí.

2. Pokud formule obsahuje emoční výrazy označené emoční třídou, nahradíme {NEG2}, {NEG}, {NEU} , {POS}, {POS2} příslušným emočním koeficientem slova.

3. Pokud formule obsahuje negace následujícího slova {!}, pak emoční koeficient následujícího slova bude vynásoben číslem -1.

4. Pokud formule obsahuje multiplikátory {Mn}, je nalezen nejbližší vyjádřený sentiment a uplatní se koeficient vynásobením emočního koeficientu slova. Často jsou multiplikátory přímo vedle slov vyjadřujících sentiment, ale pokud tomu takto není, hledá se dle vzdálenosti a hodnot $\{n\}$, kde n je počet za sebou jdoucích neurčených slov.

5. Celkový emoční koeficient věty se stanovuje jednoduchým zprůměrováním dílčích emočních koeficientů slov ovlivněných negací či multiplikátory.

6. Pokud formule obsahuje negaci celé věty {!V} dojde k vynásobení výsledného emočního sentimentu číslem -1.

Postup výpočtu koeficientu emoční třídy a její stanovení je demonstrováno na příkladu 5.7.

Příklad 5.7

Ukázku vyhodnocení provedeme na formuli z příkladu 5.3, kde byla věta „Fotoaparát není příliš dobrý na focení v tmavých místnostech.“ převedena na formuli „{PA123} {S}{!V} {M0.5} {POS} {5}“. Budeme postupně aplikovat kroky z postupu uvedeného výše.

1. Formule neobsahuje vulgarismy {VUL}, takže ohodnocení může pokračovat.
2. Formule obsahuje emoční výrazy {POS}, který je nahrazen příslušným emočním koeficientem slova, tedy označením „{SE1}“. Jednička vyjadřuje koeficient sentimentu.
3. Formule neobsahuje negace následujícího slova {!}.
4. Formule obsahuje multiplikátor {M0.5}, který musí být použit na nejbližší slovo vyjadřující sentiment, ten je již přepsaný na číslo, kterým je hned následující {POS} respektive číslo 1. Hodnota se tedy změní na 0.5. Po tomto kroku upravená formule původní věty vypadá takto: „{PA123} {S}{!V} 0.5 {5}“.
5. Jelikož formule obsahuje pouze jedinou hodnotu a to 0.5, jedná se tak také o výslednou zprůměrovanou hodnotu.
6. Formule obsahuje negaci celé věty {!V}, takže dojde k vynásobení výsledného emočního sentimentu číslem -1 a výsledný sentiment je tedy -0.5.

Pokud se podíváme do tabulky emočních tříd a jejich koeficientů 13, jedná se v případě výsledné hodnoty -0.5 o negativní sentiment.

Výsledkem analýzy sentimentu je: Věta „Fotoaparát není příliš dobrý na focení v tmavých místnostech.“ se nachází v emoční třídě negative s koeficientem -0.5. ■

Pokud provádíme analýzu sentimentu s ohledem na jednotlivé aspekty, je postup rozšířen o další kroky. V nich se řeší přítomnost aspektů v samotné větě či sdílení aspektu z věty předešlé. Pokud dojde k dědění aspektu, také je do formule před aspekt z předešlé věty vložena automaticky vzdálenost deset slov. Na začátek formule se tedy přidá „{A#ID}{10}“. Každý aspekt může být sdílen pouze v následující větě. Potenciální aspekty {PA#ID} není možné sdílet.

Při vyhodnocování se pro každý vyjádřený sentiment hledá nejbližší aspekt. Pokud se ve větě nevyskytuje, je výsledkem pouze sentiment celé věty. Pokud by ale v uvedeném příkladu 5.7 bylo slovo „focení“ aspektem, pak by se jednalo o jediný a zároveň

nejbližší aspekt ve větě. Výsledkem by bylo: Věta „Fotoaparát není příliš dobrý na focení v tmavých místnostech.“ obsahuje aspekt „focení“, který se nachází v emoční třídě negative s koeficientem -0.5.

Může nastat také situace, kdy se v jedné větě nachází více slov vyjadřujících sentiment a zároveň také více aspektů. Na příkladu 5.8 je vidět, jak je analýza sentimentu v takové chvíli řešena.

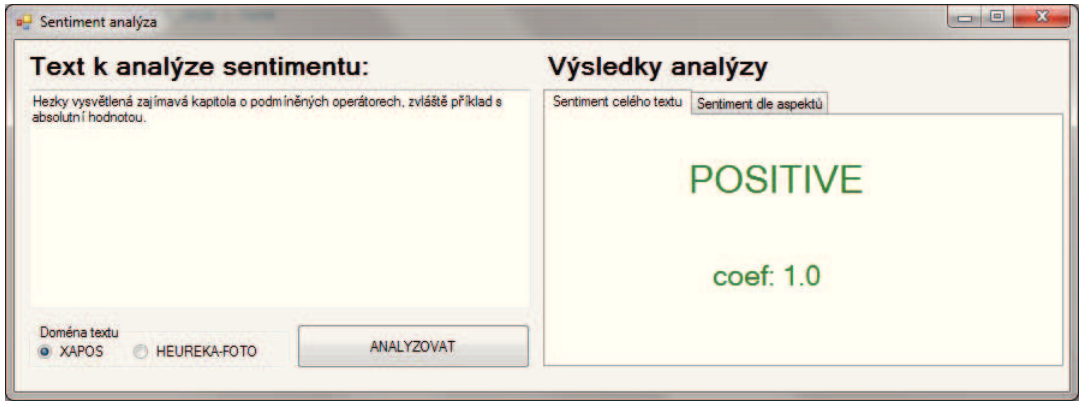
Příklad 5.8
Ukázku vyhodnocení provedeme na větě „Výborná čočka v kombinaci se slušným zoomem“, která je převedena na formuli ”{POS2}{A12} {3} {POS2}{A23}”. Po aplikování standardních kroků pro vyhodnocení získáme ohodnocenou formuli „ 2 {A12} {3} 1 {A23}“. Pro každý koeficient emoční třídy hledáme aspekt, pro který je daný sentiment vyjádřen. V případě hodnoty 2, tedy slova „Výborná“ je nejbližším aspektem slovo „čočka“ s ID 12. Pro emočním koeficient 1 respektive slovo „slušným“ je vybrán aspekt zoom s ID 23. ■

5.4.6 Prezentace a uložení výsledků

Texty, příspěvky nebo recenze jsou analyzovány zcela samostatně, takže na konci analýzy sentimentu záleží na zvolené aplikaci, jak s vyhodnocenými daty naloží. Výsledky analýzy je tak možné zapisovat do databáze, prezentovat uživateli nebo třeba odesílat na vzdálený server.

V případě analýzy sentimentu na úrovni aspektu je vhodné zapisovat jednotlivé záznamy z analýzy do tabulky k tomuto účelu vytvořené. Uložení ID aspektu, ID analyzovaného textu společně s koeficientem emoční třídy je pro základní použití dostačující. Obrázek 9 zobrazuje výsledek analýzy v GUI testovací aplikace.

Může být výhodné do tabulky ukládat rovněž konkrétní větu, která ukládaný sentiment obsahuje. Takto uložený text významně ulehčí pokročilejší analýzu sentimentu, kdy může být například zjišťováno, co konkrétně se uživatelům v souvislosti s daným aspektem nelíbí. Tato analýza je pro automatický způsob až příliš podrobná, vyžaduje cit i znalost problematiky. Je vhodné to uživatelům co nejvíce ulehčit uložením věty s daným aspektem a emoční třídou. Příklad 5.9 důležitost uchování věty s daným aspektem a emoční třídou jasně dokazuje.



Obrázek 9: GUI Aplikace pro analýzu sentimentu

Příklad 5.9
Pokud by chtěl výrobce fotografií analyzovat negativní sentiment v recenzích k produktu pro jednotlivé aspekty, bude pro něj informace o tom, že deset lidí hodnotí fotografie negativně, poměrně významná. V případě, že bude chtít analyzovat všech deset výskytů negativní emoční třídy, může si jednoduše zobrazit věty, ze kterých bude schopen zjistit, co konkrétně se těmto zákazníkům na fotografiích nelíbí. Nebude nucen zkoumat celé recenze.

Mohou to být třeba věty: „Fotografie jsou velice nekvalitní“ či „Bohužel mám všechny fotografie rozmazané:(“. Výrobce v případě uložení těchto částí do databáze není nucen vyhledávat informace v rozsáhlém textu. ■

5.4.7 Návrh nových aspektů

Je důležité mít slovník aspektů a jejich synonym co nejobsáhlejší. Nejčastěji se slovník aspektů plní manuálně, je ale možné využít také automatického navrhování aspektů. Metoda pro hledání nových aspektů byla popsána v pramenu [22], ale je vhodná spíše pro použití v algoritmech analýzy sentimentu využívajících strojové učení. Pro využití při slovníkové metodě analýzy sentimentu jí bylo nutné významně modifikovat.

Potenciální aspekt je určen dle pozice ve větě vyjadřující sentiment. Je možné definovat dvě pozice ve větě, kde je výskyt aspektu nejpravděpodobnější. V prvním případě se jedná o unigram vyskytující se za slovem patřícím do jedné z emočních tříd. Kandidátem na aspekt je rovněž slovo vyskytující se před slovesem.

Každý výskyt potenciálního aspektu je zaznamenán do databáze a později může proběhnout jeho schválení či zamítnutí. U potenciálních aspektů je evidován také počet výskytů, který definování nového aspektu významně ulehčí.

5.5 Výsledky experimentu

Analýzu sentimentu bylo možné úspěšně otestovat ve dvou různých doménách a to v e-learningovém systému XAPOS[16] [27] a na recenzích umístěných na serveru Heureka[17]. V obou případech byly výsledky převedeny na cenné informace, které mají skutečný význam.

Ze systému XAPOS jsme získali důležité informace o tom, jak jsou studenti spokojeni s obsahem a s jednotlivými aspekty celého systému. Tabulka 14 zobrazuje několik aspektů a jejich výskyt v emočních třídách. Při aplikaci na doménu XAPOSu je důležité brát v úvahu fakt, že každý vzdělávací objekt, ke kterému se komentář váže, je analyzován samostatně. Například pro aspekt tabulka v komentářích k jednomu určitému vzdělávacímu objektu bylo pouze šest výskytů sentimentu v emoční třídě negative. Při podrobnější analýze vět bylo zjištěno, že se v daném objektu tabulka nezobrazuje. Analýza sentimentu tak přispěla ke zkvalitnění obsahu.

aspekt	**výskyt emočních tříd**
tabulka	11 negative, 4 neutral, 6 positive
rekurze	1 negative, 1 positive

Tabulka 14: Emoční třídy vyhodnocené pro aspekty

Měření úspěšnosti analýzy sentimentu je založeno na porovnávání emočních tříd určených automatizovaným algoritmem vůči manuálnímu určení u náhodně vybraných komentářů. Pro srovnání s jinými algoritmy je porovnáváno pouze určování sentimentu celé věty, jelikož pouze v takovém případě jsme schopni výsledky jednoznačně porovnat.

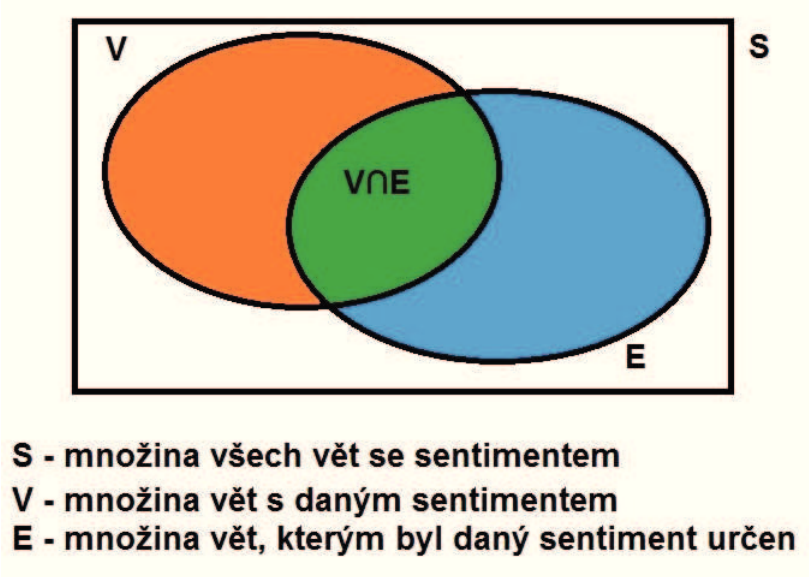
Pro definování úspěšnosti analýzy sentimentu je využito *f-skóre*. To je zvoleno z důvodu možného srovnání s metodou analýzy sentimentu v česky psaných textech prezentovanou v diplomové práci Radka Července [2], který však využívá metodu strojového učení a nezohledňuje aspekty nalezené v textu.

[16] http://arg.vsb.cz/XAPOS/, authors: Zdeněk Velart, Petr Šaloun
[17] http://www.heureka.cz

Úspěšnost algoritmu pro analýzu sentimentu je těžce měřitelná. Hlavním důvodem je fakt, že pravděpodobnost shody dvou lidí při určování sentimentu je 80 %. V pramenu [11] je také uvedeno, že úspěšnost těch nejlepších algoritmů se pak pohybuje mezi 70 % a 80 %. Protože výsledky automatické analýzy sentimentu byly porovnávány s mým subjektivním vyhodnocením sentimentu daného textu, jsou výsledky úspěšnosti velmi vysoké. Průměrná hodnota *f-skóre* odpovídá hodnotě 0.79, přitom maximum je 1. Pokud by ale výsledky automatické analýzy byly porovnávany s manuálním určením jiné osoby, výsledky úspěšnosti by byly jiné.

5.5.1 F-skóre

Úspěšnost správné klasifikace emoční třídy a celé sentiment analýzy lze určit pomocí *f-skóre*. Význam a použití *f-skóre* (f-measurement) je vysvětlen v pramenu [29]. Pro stanovení *f-skóre* se využívá dílčích výpočtů *přesnosti* (precision) a *výtěžnosti* (recall).



Obrázek 10: Prostor vět se sentimentem

Definice 5.4 *Přesnost klasifikace je schopnost algoritmu zařadit do dané emoční třídy pouze relevantní věty. Je vyjádřena poměrem počtu vět správně zařazených do vybrané emoční třídy ku počtu všech vět algoritmem zařazených v dané emoční třídě.*

$$P = \frac{|V \bigcap E|}{|E|}$$

Definice množin V a E je uvedena na obrázku 10.

Definice 5.5 *Výtěžnost klasifikace je schopnost algoritmu zařadit větu do správné emoční třídy. Je vyjádřena poměrem počtu vět správně zařazených do vybrané emoční třídy vzhledem k počtu vět, které měly být do dané emoční třídy správně zařazeny.*

$$R = \frac{|V \bigcap E|}{|V|}$$

Definice množin V a E je uvedena na obrázku 10.

Definice 5.6 *F-skóre je kompromisní hodnotou mezi přesností a výtěžností.*

$$\mathrm{f-skóre} = 2 \times \frac{P \times R}{P + R}$$

Jak se ukazatele úspěšnosti aplikovaného algoritmu *přesnost*, *výtěžnost* a *f-skóre* pohybovaly je uvedeno v tabulce 15.

emoční třída	**f-skóre [-]**
vulgar	1
negative2	0.82
negative	0.72
neutral	0.82
possitive	0.74
possitive2	0.81

Tabulka 15: Úspěch analýzy sentimentu vyjádřený pomocí f-skóre

5.6 Problémy slovníkové metody

Analýza sentimentu založená na slovníkové metodě má však několik obecně známých úskalí a problémů. Některé z nich jsou popsány v knize [12]. Při navrhování algoritmu ale tyto slabé stránky byly brány v úvahu. Algoritmus se snaží jimi způsobenou nepřesnost v určování emoční třídy eliminovat.

5.6.1 Cizí slova, zkratky a slangové výrazy

Často je při tvorbě slovníků pro analýzu sentimentu využito slovníků jako například WordNet[18]. Tyto slovníky významně ulehčují nalezení slov vyjadřujících emoce či míru emocí, ale obvykle obsahují jen spisovné výrazy. Při analýze sentimentu tedy hrozí, že některé *n-gramy* vyjadřující emoce nebudou rozpoznány. Může se jednat o slang, zkratky či výrazy cizího jazyka.

Díky použité poloautomatické metodě tvorby slovníků je možné jejich doplnění o další výrazy. Často se dnes můžeme na českých diskusích setkat například s výrazem „thx“, který je zkrácenou formou anglického „Thank you.“. V tomto případě se tedy jedná o zkratku cizího výrazu vyjadřující pozitivní sentiment. V českých slovnících bychom ale „thx“ hledali marně. Při plnění využitých slovníků bylo doplněno několik často používaných výrazů slangu, zkrácených forem a cizích slov. Obdobný problém lze očekávat v případě vulgarismů.

5.6.2 Sarkasmus

V česky psaných textech se s užitím sarkasmu ve větší míře nesetkáváme, nicméně stále je to jeden z problémů sentiment analýzy. Například věta „Opravdu skvělá stránka, zobrazil se mi jen nadpis!“ by byla chybně kvalifikována do třídy positive. První část věty vyjadřuje silně pozitivní emoce, druhá část věty je analyzována jako lehce negativní díky samostatnému použití bez výskytu n-gramu vyjadřujícího sentiment. Člověkem, který rozpozná sarkasmus, je ale věta vyhodnocena jako negativní. Velký pokrok v oblasti rozeznání sarkasmu v anglicky psaných příspěvcích s f-skóre okolo 0.8 je popsán v [32].

5.6.3 Neznalost kontextu a souvislostí

Některá slova mají v určitém kontextu význam předmětný a nemají za úkol vyjadřovat emoce. Klasickými příklady jsou například „Škoda“ nebo „bomba“. Pokud se tato slova vyskytnou na diskusi o autech respektive válečné historii, je dost velká pravděpodobnost, že jsou použita v souvislosti s věcným předmětem a nikoli sentimentem. Z tohoto důvodu je vhodné pro každou specifickou doménu textů použité slovníky pro rozpoznání n-gramů upravit.

[18] wordnet.princeton.edu/

5.6.4 Víceznačná slova

V češtině existuje řada slov, jejichž význam se zásadně mění v závislosti na dalších výrazech použitých ve větě. Některá navíc mohou vyjadřovat při různých použitích naprosto opačný sentiment. Tuto situaci vhodně znázorňuje příklad 5.10.

Příklad 5.10
Jedním ze slov, které mohou mít naprosto opačný význam při analýze sentimentu je unigram „pohodlný". Pokud je použito ve větě „Máš moc pohodlný gauč", vyjadřuje pozitivní sentiment. Jestli se ale tento výraz vyskytne ve větě „Jsi pohodlný člověk!" měl by vyjadřovat emoční třídu negative. V tomto případě algoritmus nemá možnost rozpoznat, o který význam se jedná. ■

Jedním z možných řešení je vyřadit taková slova ze slovníku, tím podaří vyhnout možnému špatnému přiřazení emoční třídy, na druhou stranu ale nedojde ani k přiřazení té správné. Komplikovanější možností je přidat sadu podmínek na výskyt určitých slov v okolí. Díky rozmanitosti češtiny ale ani jedna z možností není jednoznačně úspěšným řešením. Jinak tomu bylo v případě víceznačného slova „boží" v příkladě 5.4.

5.6.5 Jazyková čistota

Na kvalitu analýzy sentimentu má velký vliv užitá čistota jazykových výrazů. Pokud jednotlivé unigramy obsahují gramatické chyby či překlepy, není výraz nalezen ve slovnících a jeho význam není rozpoznán. Pro češtinu již existuje řada aplikací či doplňků prohlížeče, které kontrolu pravopisu (spell chacking) umožňují. Samotné zařazení kontroly pravopisu do algoritmu pro analýzu sentimentu je jedna z možností. Různé přístupy pro kontrolu pravopisu v česky psaných textech jsou diskutovány v [30].

6 Zhodnocení a závěr

Cílem této práce bylo analyzovat chování uživatelů v adaptivním webu a zformulovat hypotézu o zkvalitnění navigace na základě znalosti významu a kategorie obsahu.

Pro experiment byl využit e-learningový systém XAPOS fungující v doméně VŠB-TU Ostrava. Do systému bylo implementováno přidávání komentářů, anotací, hodnocení anotací a možnost volby stupnice hodnocení anotací. V rámci experimentu systém XAPOS použily dvě stovky studentů a bylo získáno velké množství dat k analýze. Byl aplikován postup pro rozlišení kvality práce studentů v systému, z něhož vyplynulo, že u 88 studentů z 200 existuje podezření, že se v systému nechovali správně a jejich data nemusí být pro další analýzu dostatečně kvalitní, podrobněji viz kapitola 3.3.1.

Studenti měli možnost za svou aktivitu získat bonusové body. Při analýze výsledků studentů v systému XAPOS a ve studijním předmětu bylo zjištěno, že studenti, kteří v XAPOSu pracovali lépe a byli aktivnější, získali více bodů také ve studijním předmětu.

Byla analyzována doba strávená uživateli na jednotlivých stránkách a zjištěn počet návratů na jednotlivé stránky. Na některých stránkách uživatelé strávili průměrně jen několik málo desítek sekund. Tyto stránky jsou navrženy na kontrolu obsahu. Naopak v systému XAPOS jsou také stránky, na které se uživatelé často vracejí. Takové chování uživatelů může naznačovat, že navigace nefunguje zcela správně.

S využitím vizualizace bylo předvedeno, že lze velice přehledně sledovat, jak se uživatelé v systému pohybují. Vizualizace v tomto případě pomáhá více porozumět uživatelům a výsledky její analýzy mohou vést k úpravě navigace i obsahu.

Navigace v systému XAPOS využívá prostoru konceptů a množiny dosažených znalostí jednotlivých uživatelů. Koncepty popisují obsah jednotlivých výukových objektů. Jedna z možností, jak odhalit nepřesnost nastavených konceptů a jejich vazeb, je analýza sentimentu komentářů k danému konceptu a výukovému objektu. Pokud koncept a stránka samotná budou uživateli špatně hodnoceny, jedná se o indikaci možného problému stránky. To může při podrobnější analýze vést k úpravě obsahu nebo navigace.

Hlavním přínosem této práce je navržení, realizace a ověření metody pro analýzu sentimentu v česky psaných textech. Uživatelé v systému zanechali více než 1400 komentářů, z nichž velká část vyjadřuje názor. Tento názor má smysl automaticky analyzovat a zjistit jeho sentiment. Výsledek analýzy sentimentu je zpětnou vazbou a může být podnětem ke změnám v systému z hlediska navigace i obsahu.

Analýzou sentimentu česky psaných textů se mnoho prací nezabývá. Bylo možné se inspirovat různými přístupy aplikovanými na texty v angličtině. Na základě získaných znalostí byla zvolena metoda analýzy sentimentu založena na slovnících. Ta byla doplněna o prvky používané v jiných metodách řešících NLP problémy. Pomocí nich se podařilo dosáhnout lepších výsledků.

Čeština je specifickým jazykem, a z toho důvodu byl vytvořen *slovník výjimek*, který se stal klíčovou součástí námi vytvořené metody pro analýzu sentimentu. Díky němu je při určování sentimentu důležitý nejen výraz samotný, ale také výrazy použité v celé analyzované větě. Úspěšnost algoritmu se podařilo zlepšit dvojitým testováním, kdy je význam jednotlivých výrazů vyhledáván ve slovnících před i po převedení na základní tvar. Při návrhu přístupu k analýze emocí z textu jsme se pokusili eliminovat nevýhody vyplývající z použití slovníkového přístupu. Jedná se například o sarkasmus, použití cizích slov nebo dvojznačnost výrazů.

Součástí metody je také rozpoznávání *aspektů* nebo-li předmětů hodnocení. Schopnost rozpoznat předmět názoru je velice cenná a může při pokročilém vyhodnocování sentimentu výrazně pomoci. Například díky definování aspektu „tabulka“ a následné analýze sentimentu komentářů bylo zjištěno, že na jedné ze stránek v XAPOSu se uživatelům nezobrazuje tabulka. Součástí řešení je také metoda pro automatické navrhování nových aspektů. Výsledky z testování ale ukázaly, že správnost navržení nového aspektu je velice nízká.

Metoda byla použita pro analýzu komentářů v systému XAPOS a na recenze produktů na stránkách heureka.cz[19]. Průběžné testování analýzy sentimentu obsahu z obou zmiňovaných domén bylo pro definování konečné podoby metody velice prospěšné. Hodnota f-skóre, která se používá k vyjádření úspěšnosti analýzy, je rovna hodnotě 0.79 z 1. To značí poměrně vysokou úspěšnost analýzy sentimentu. Hodnota f-skóre je v našem případě srovnatelná s jinými běžně používanými algoritmy.

Podstatná část této diplomové práce byla shrnuta v článku přijatém na konferenci „Interdisciplinary Symposium on Complex Systems - ISCS2013“ [20], konanou ve dnech 10. - 13. září 2013 v Praze, kde bude prezentován odborné veřejnosti. Článek je přiložen v příloze A a bude publikován v knize vydané nakladatelstvím Springer.

[19] http://www.heureka.cz/

[20] https://sites.google.com/site/complexsystems2013/home

6.1 Budoucí práce

Při analýze sentimentu se podařilo dosáhnout poměrně vysoké úspěšnosti, navržená metoda však může být dále rozšiřována a zlepšována. Přesnost určování je možné zvýšit obohacením všech slovníků. Nejvíce klíčový je slovník výjimek, který řeší kontext dané věty a ne jen jednotlivá slova.

Při testování metody analýzy sentimentu bylo zjištěno, že se nepodařilo řadu slov určit vinou překlepů a gramatických chyb. Neurčená slova mají na úspěšné rozpoznání emocí podstatný vliv. Možným řešením by v tomto případě mohlo být doplnění metody o automatickou kontrolu pravopisu. Prostor pro možné úpravy je rovněž v přístupu pro rozpoznání nových aspektů, kde by mohlo být využito definování přísnějších podmínek pro navrhování nebo nový přístup. Zajímavou možností rozšíření je rozpoznávání sarkasmu v česky psaném textu.

7 Reference

[1] Brusilovsky T., Klobsa A., Nejdl W.: The Adaptive Web, 2007. pp. 766.

[2] Cervenec R., Burget R.: Identifying Expression of Emotions in Czech Text Using Semantic Relations for Dimension Reduction, Elektrorevue, Vol. 2 No. 3. 2011. pp. 16-21.

[3] Feldman R., Rozenfeld B.,Y. Breakstone M.: A Hybrid Approach to Sentiment Analysis of Stocks. In ISCOL 2010, Israeli Seminar on Computational Linguistics. 2010. pp. 1-16.

[4] Kamps, J., Marx, M., Mokken, R.J., de Rijke, M.: Using WordNet to measure semantic orientation of adjectives. LREC. 2004. pp. 1-4.

[5] Prabowo R., Thelwall M.: Sentiment Analysis - A Combined Approach, School of Computing and Information Technology University of Wolverhampton. 2009. pp. 143-157.

[6] Peter Brusilovsky and Eva Millán User Models for Adaptive Hypermedia and Adaptive Educational Systems , The Adaptive Web, LNCS 4321. 2007. pp. 3–53.

[7] Šlerka J.: O sentiment analýze bez sentimentu aneb jeden malý experiment, lupa.cz. 2011. [online: 12.4.2013]

[8] Neviarouskaya A., Prendinger H., Ishizuka M. Semantically distinct verb classes involved in sentiment analysis, IADIS International Conference Applied Computing 2009. 2009. pp. 27-34.

[9] De Bra P., Houben G.J., Wu H.: AHAM: A Dexter-based Reference Model for Adaptive Hypermedia. In Proceedings of the ACM Conference on Hypertext and Hypermedia. ACM New York. 1999. pp. 221-239.

[10] O'Reilly T., Battelle J.: Web 2.0 Five Years On, WEB2.0 SUMMIT - Special Report, O'Reilly Media, Inc., 2009. p. 15.

[11] Marinka Zitnik: Using sentiment analysis to improve business operations. ACM Crossroads 18(4). 2012. pp. 42-43.

[12] Liu, B.: Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2012. p 168.

[13] Vander T.W.: Understanding Folksonomy: Tagging that Works. Brighton England, on http://archive.dconstruct.org/2006/understandingfolksonomy. 2006. [online:15.6.2013]

[14] Indurkhya N., Damerau Fred J.: Handbook of Natural Language Processing, Second Edition. 2010. p. 704.

[15] Sanda P.: Určení základního tvaru slova, Diplomová práce na Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií ústav komunikací. 2011. p. 68.

[16] Dolamic L., Savoy J.: Stemming Approaches for East European Languages, Advances in Multilingual and Multimodal Information Retrieval Lecture Notes in Computer Science Volume 5152. 2008. pp. 37-44.

[17] Lovins J. B.: Development of a Stemming Algorithm, Mechanical Translation and Computational Linguistics, 1968, pp. 1-10.

[18] Z. Velart: Adaptivní personalizovaná navigace řízená metadaty, Disertační práce na Vysoká škola báňská Technická univerzita Ostrava, 2011, p 104.

[19] Bespalov D. Sentiment Classification Based on Supervised Latent n-gram Analysis, the 20th ACM Conference on Information and Knowledge Management. 2011. p. 28.

[20] Hartmann T., Klenk S., Burkovski A., Heidemann G.: Sentiment Detection with Character n-Grams, Proc. 2011 Int. Conf. on Data Mining (DMIN'11). 2011. pp. 364-368.

[21] Fellbaum C.: Wordnet: An Electronic Lexical Database. MIT Press, Cambridge, MA. 1998. p. 423.

[22] Wu Y., Zhang Q., Huang X., WuPhrase L.: Dependency Parsing for Opinion Mining, Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. 2009. pp. 1533–1541.

[23] Husaini M., Koçyigit A., Tapucu D., Yanikoglu B., Saygin Y.: An aspect-lexicon creation and evaluation tool for sentiment analysis researchers. In Proceedings of the

2012 European conference on Machine Learning and Knowledge Discovery in Databases - Volume Part II (ECML PKDD'12), Peter A. Flach, Tijl Bie, and Nello Cristianini (Eds.), Vol. Part II. Springer-Verlag, Berlin, Heidelberg. 2012. pp. 804-807.

[24] Šaloun P., Velart Z., Nekula J.: Towards automated navigation over multilingual content. Springer, Studies in Computational Intelligence 418. 2013. pp. 203-229.

[25] Šaloun P., Velart Z.: Adaptive ontology-based navigation. In Proceedings of A3H: 6th International Workshop on Authoring of Adaptive and Adaptable Hypermedia. Springer. 2008. p. 4.

[26] Sujatha V.: An approach to user navigation pattern based on ant based clustering and slassification using decision trees, International Journal of advanced engineering sciences and technologies Vol No. 1, Issue No. 2. 2010. pp. 112-117.

[27] Velart Z., Nekula J., Šaloun P. Experimentální adaptivní vícejazyčný webový systém. In ITAT 2010 - Informaené Technológie Aplikácie a Teória. 2010. pp. 129-130.

[28] Doreian P., Stokman F.: Evolution of Social Networks. 2013. p. 272.

[29] Van Rijsbergen C. J.: Information Retrieval (2nd ed.). Butterworth-Heinemann, Newton, MA, USA. 1979.

[30] Bureš, S.: Kontrola pravopisu v českých textech, Diplomová práce na Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií ústav komunikací. 2011. p 48.

[31] Kashyap V., Bussler Ch., Moran M.: The Semantic Web Semantics for Data and Services on the Web, 2008. p. 414.

[32] Davidov D., Tsur O., Rappoport A.: Semi-supervised recognition of sarcastic sentences in Twitter and Amazon. In Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL '10). Association for Computational Linguistics, Stroudsburg, PA, USA. 2010. pp. 107-116.

[33] Yared P.: Why sentiment analysis is the future of ad optimization, venturebeat.com. 2011. [online: 15.7.2013]

[34] Sosnovsky, S., Brusilovsky, P., H. Lee, D., Zadorozhny, V., Xin Zhou: Re-assessing the Value of Adaptive Navigation Support in E-Learning Context. In: Proceedings

of the 5th international conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH '08). 2008. pp. 193-203.

[35] M. Briš: Analýza správania sa používateľa webu, Slovenská technická univerzita v Bratislavě, Fakulta infromatiky a komunikačních technologií. 2013. p. 50.

[36] Bastian, M.; Heymann, S.; Jacomy, M.: Gephi: An Open Source Software for Exploring and Manipulating Networks. International AAAI Conference on Weblogs and Social Media, North America, mar. 2009 (ICWSM09). 2009. p. 2.

A Příloha A

Článek na téma *Sentiment Analysis* přijatý na konferenci „Interdisciplinary Symposium on Complex Systems - ISCS2013“ [21] konanou ve dnech 10. - 13. září 2013 v Praze, kde bude prezentován odborné veřejnosti. Článek bude publikován v knize vydané nakladatelstvím Springer.

[21] https://sites.google.com/site/complexsystems2013/home

Sentiment analysis in complex adaptive systems

Petr Šaloun, Ivan Zelinka, and Martin Hruzik

VSB-Technical University of Ostrava, 17. listopadu 15, 70833 Ostrava, Czech Republic,
petr.saloun@vsb.cz, me@martinhruzik.cz, ivan.zelinka@vsb.cz

Abstract. The aim of this work is to present a new algorithm for the evaluation of sentiment in Czech language texts. The algorithm is based on a new dictionary and uses n-gram searching. For the creation of the dictionary, it was important to use language specific phrases and exceptions, which can completely change the final evaluation of a sentiment. The solution also includes automatic search for a new subjects (aspects) of evaluation and also searching for new words determining sentiment. A similar algorithm can also be applied to other languages. The work emphasizes the transformation of the acquired data into valuable information. Our experiment is realized in the experimental adaptive web system in e-learning content domain and in eShop domain. The success and benefits of the algorithm are also discussed in this text.

Keywords: sentiment analysis, opinion mining, lemmatization, aspects mining, Wordnet

1 Introduction

The Internet is constantly replenished with new users who generate more and more content. The content conceals information of high value, and many companies, scientists, prefessionals and experts are trying to figure out how to obtain and use this valuable information. Subjective texts, which are important for their sentiment, are a part of this.

Sentiment analysis (opinion mining) is currently one of the most discussed topics. In the past, there have been several startups and a number of methods created with the same goal - get the sentiment from text. These methods are used for example in financial markets, where sentiment analysis helps with stock trading.

The most common application of sentiment analysis is in the area of consumer reviews of products and services. Opinion mining helps producers to find out what the customers think about their products and determine what they like and what they complain about. This information can help manufacturers to further develop a product, and subsequently, increase sales. Customers on the other hand, can see what other people refer to as an advantage or disadvantage and then decide whether to buy the product or not. There are numerous news items, articles, blogs, tweets etc. which are analyzed [1].

It is possible to find many different ways for doing sentiment analysis of English texts. Some of these methods have been already implemented for Chinese and Spanish texts. The other languages usually use the common solution - first, the text is converted into English, and then the analysis of sentiment is done. Each language has its own peculiarities, so the method that involves a translation is not always successful.

This text is focused on the procedure of sentiment analysis in adaptive systems based on Slavic languages and Czech language primarily. We have used my experience from e-shops and native knowledge of the Czech language to design the algorithm. In addition to analyzing data about specific products, the algorithm was also succesfully applied in the domain of e-learning, where we gathered students' feedback. Adaptive web systems in education domain are described in [16]. We were able to improve content, user interface and usability of the system based on the result of the sentiment analysis.

2 Related work

The impulse for application sentiment analysis was the change of web standards - WEB 2.0. Since 2004, the world of the internet is not just about static websites, but users are already actively involved in the creation of content, and websites are full of interactive elements. The internet is full of subjective texts that can be further processed and analyzed for sentiment to gain valuable information [7].

Sentiment analysis gathers emotions of the author of the text. In its simplest form, the sentiment distinguishes positive and negative emotions, but there are also algorithms that are able to recognize fear, anger and other human emotions. Subjective texts, with identifiable emotions, are analyzed more frequently. There is also sentiment analysis of objective text based on the facts - for example monitoring of the financial market.

In general, sentiment analysis has been investigated mainly at three levels [2]:

Document Level: Result is the identification of positive, negative or neutral sentimnetu for the entire document. It is assumed, that each document contains text related to only one entity.

Sentence level: At this level, the task is to determine whether each sentence expresses positive, negative, or neutral opinion.

Entity and Aspect level: The document level and the sentence level analyzes do not exactly determine what people liked and disliked. Sentiment analysis at the aspect level is based on the rule that each opinion consists of sentiment (positive, negative) and objects (target of opinion).

One of the main conditions for determining the correct sentiment by the algorithm based on key words is quality of the lexicon. Lexicon contains sentiment words, also called opinion words, and multiplication words (almost, so, really etc.).

It is important to mention, that the sentiment analysis is a NLP (Natural Language Processing) problem. Success of the analysis also depends on quality of NLP for the chosen language, which is Czech in this case. There are many NLP methods and quality dictionaries for English. However, finding the right NLP techniques and vocabulary for other languages is much more complicated [3,5].

Sentiment analysis in the Czech text is still a relatively unexplored area. When we were looking for the existing solutions, we found a work from 2011, where instead of dictionary, machine learning is used [8]. One of the most popular methods of machine learning is SVM (Support Vector Machine) [17]. The success of this method depends directly on the quality of the training set, which may be specific for different areas of entity [9].

There is also a hybrid method that combines the benefits of dictionary and machine learning approaches [10]. This article focuses on sentiment analysis of aspects based on the use of the new lexicon.

3 Sentiment analysis

There are many emotions, but we only distinguish between positive, neutral and negative emotions for the resulting sentiment. We also divide the power of emotion into two groups - normal and strong. Special group are sentences which contain vulgar phrases and a negative sentiment is used. Table 1 shows all the defined classes of sentiment and also an example.

Table 1. Emotion classes

emotion class	**description**	**example**
vulgar	Vulgar and negative emot.	"Buying this shit was a huge mistake!"
negative2	Srong negative emotions	"I hate this camera, I am so angry about that."
negative	Normal negative emotions	"This camera is not good."
neutral	Neutral emotions	"I found some issues but it is " not a bad camera."
positive	Normal positive emotions	"I can recommend this camera."
positive2	Strong positive emotions	"I love this camera I am so happy!"

In the first step, the text is converted to the lower case format, and the system detects whether diacritics is used. If the diacritics is used, the system will use the original comparison in the next steps. Then the text is split to sentences using special characters and stopwords. In this context, sentence means a unit of sentiment analysis.

Each unit (sentence) is going through lemmatization - the words and expressions are converted to the basic word form by using the rules for inflection or the special dictionary. This problem is language specific, so the author's sense for the Czech language was used, and also the methods from [12]. Once we have a sentence in basic form, we can define n-grams (unigrams, bigrams, trigrams) [13].

With n-grams, accuracy of the algorithm is improved, because a comparison is done with words and their dependencies, not just each one separate word.

The main phase of the Sentiment Analysis is Formula Transformation. In this step, the application transforming text into stream of symbols. All n-grams are processed with the exception of dictionary, keywords dictionary, aspect list and emoticon list. The output is a formula, which is used for the final classification of the emotion class and relation with aspect. Table 2 describes the sentence-to-formula transformation.

Table 2. Formula transformation

input	**output**
"This camera has really amazing zoom."	{3} {M1.8} {POS2} {A123}
Explanation: {3} - three unrecognized words in row, "really" - {M1.8}(multiple word with cofficient), "amazing" - {POS2}(StrongPositive), "zoom" - {A123} aspect with ID 123	

The last part of the sentiment analysis is sentiment classification. In this step, evaluation and determination of the final sentiment of the text or aspect is done. For the example above is the result: Aspect - zoom, Sentiment - strong positive, Points +4.

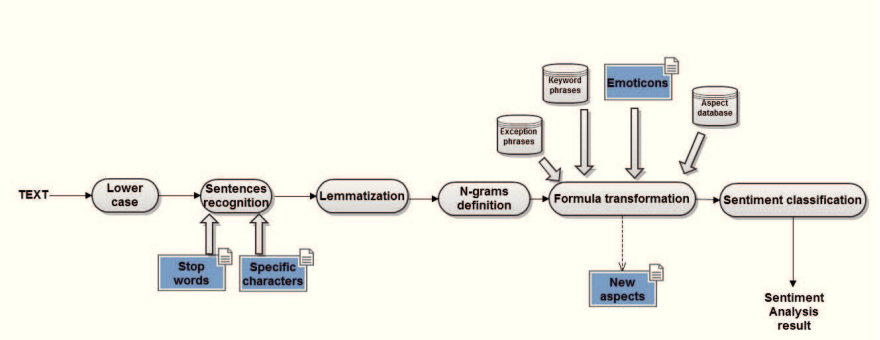


Fig. 1. The proposed system for sentiment analysis

3.1 Lexicon based acquisition

A good lexicon is the base of all methods of the lexicon sentiment analysis. The dictionary contains several types of phrases that have the main significance in the sentiment analysis [6], see Table 3.

Table 3. Keywords types

type	description
Sentiment phrases	Phrases which identify positive or negative sentiment
Multiple phrases	Phrases which identify multiplication of the sentiment
Aspects	Attributes/properties of the subject
Exceptions	Phrases with special meaning

For the creation of the dictionary, several automatic and manual methods can be used. One of the interesting automatic methods is based on the Wordnet lexicon [15], which contains the relations between words and is described in [14]. Since Czech is a very varied language, a semi-automatic method was used for generating the dictionary. This provides control over the phrases. At the beginning, we defined several different phrases. The application automaticly gathers synonyms and semantically similar phrases in freely available dictionaries of synonyms of the Czech language. These dictionaries are based on Thesaurus [1]. Similar phrases were automatically proposed with the level of the sentiment related to the phrase or the same level [4].

Because lexicon is a very significant for our method , the new phrases were checked and others were added manually. With a sense for the Czech dictionary, it was possible to extend the exceptional phrases, which include the more complicated phrases. There are also special phrases which can significantly affect the final sentiment.

3.2 Aspect based analysis

Aspect-based sentiment analysis is a research problem that focuses on the recognition of all sentiment expressions within a given document and the aspects to which they refer [1]. Aspect-based sentiment analysis helps gather the right information from the text. It is good to know whether the customer feels positively or negatively about a product, but by using aspects, we are able to determine what exactly is good or bad about a particular product. This information is obviously valuable.

For a definition of aspects in the e-learning domain, we used two methods. One of them is that the aspect is a key word from the ontology of the learning content text. The second is that aspects are defined manually for the entire system, for example: font, code, navigation panel. There is also a possibility to use implicit phrases, which are stored as aspect synonyms. Example of implicit aspect phrase is "This camera is too heavy", where we can recognize, that the author speak negatively about the camera weight. The word "heavy" is saved in the database of exception phrases and has a relation with aspect and also negative sentiment.

Aspect list can be extended also by using the automatic method. When the sentiment is recognized, the algorithm attempts to find a subject of the

[1] htttp://thesaurus.com/

sentiment's words. The subject is stored in a special database. If the subjects repeat in the text often, it becomes a candidate for a new aspect. For example, this way we extended aspects list of the e-learning system by adding the aspect "table" [11].

Fig. 2. Aspects of cell phone

We can store a part of the analyzed text into the same database where the results of sentiment analysis are stored. When we also save a relation to the aspect and to the sentiment class we have a quality information. With this database is easy to find all comments or part of analyzed text for the specific sentiment and aspect.

For example we can get a list of sentences with strongly negative sentiment about the camera zoom. In these sentences can be found what exactly is wrong with the camera zoom. Result of the next analysis is what needs to be done for the customers satisfaction. But this analysis is based on small details in the text so manual human sentiment analysis is the best next step.

4 Experiment results

As our experimenting environment we used elearning system XAPOS[2][18]. Almost 200 of students add to the system 1473 comments. We defined more than 20 aspects and other was set by ontology. After the sentiment analysis algorithm was applied, we got interesting picture about the system. In table below you can see a part of the output form the sentiment analysis of the student comments.

[2] http://arg.vsb.cz/XAPOS/, authors: Zdenek Velart, Petr Šaloun

Table 4. Part of the output of XAPOS analysis

aspect	**sentiment analysis**
Example	4 negative2, 6 negative, 1 positive2
Table	11 negative, 4 neutral, 6 positive
Text	2 negative2, 4 negative, 5 neutral, 31 positive, 2 positive2

From the table above we know, that we should work on navigation and change the tables. On the other hand students like text of the learning objects in the system. We also successfully applied the algorithm in eshop domain.

We compared the results of the sentiment analysis of random comments from the users in e-learning and eshop domain with our manual sentiment analysis of these comments. The succcess of the algorithm was high. F score is a performance measure that combines precision and recall and ranges between 0 (worst performance) and 100 (best performance)[19]. It is mportant to say, that there is no absolute percentage of success, because sentiment analysis is subjective for each person opinion. See Table 5. for exact numbers.:

Table 5. Success of the algorithm - F Score

type	**success** [%]
vulgar	100
negative2	82
negative	72
neutral	82
positive	74
positive2	81

5 Conclusion and future work

Our contribution is sentiment analysis system for Czech language in real working system based on the lexicon method. We have demonstrated the usability of the algorithm in the domain of e-shops, and also in domain of e-learning. Data obtained from sentiment combined with the aspect relation were transferred to valuable information, which can help improve the product or system in the educational domain.

Value of the F-score 82%, which explain the success of the algorithm, showed us, that the lexicon and whole method are not perfect and we need to work on the enhancement. For future work it is worth to consider using Wordnet or expand the lexicon by adding more n-grams and exceptions. Another option is the addition of AutoCorrection. There are also some challenges for the future work, such as irony and use of diacritics with only certain words. We believe that with these enhancements we can reach a big improvement of the algorithm for sentiment analysis of texts in the Czech language.

Acknowledgement

The following grants are acknowledged for the financial support provided for this research: Grant Agency of the Czech Republic - GACR P103/13/08195S, by the Development of human resources in research and development of latest soft computing methods and their application in practice project, reg. no. CZ.1.07/2.3.00/20.0072 funded by Operational Programme Education for Competitiveness, co-financed by ESF and state budget of the Czech Republic, and by Grant of SGS No. SP2013/114, VB - Technical University of Ostrava, Czech Republic.

References

1. R. Feldman: Techniques and Applications for Sentiment Analysis, Communications of the ACM & Vol. 56 No. 4, Pages 82-89, 2013
2. Liu, B.: Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2012.
3. A. Neviarouskaya, H. Prendinger, M. Ishizuka: Semantically distinct verb classes involved in sentiment analysis, IADIS International Conference Applied Computing 2009, Pages 27-34, 2009.
4. J.G. , Qu Y. , Wiebe J.: Computing Attitude and Affect in Text: Theory and Applications Shanahan, 2006.
5. N. Indurkhya, Fred J.: Damerau, Handbook of Natural Language Processing, Second Edition, 704 pages, 2010.
6. M. Taboada, J. Brooke, M. Tofiloski, K. Voll, M. Stede Lexicon-based methods for sentiment analysis, Association for Computational Linguistics, Vol. 37 No. 2, Pages 267-308, 2011.
7. Tim O'Reilly, John Battelle: Web 2.0 Five Years On, WEB2.0 SUMMIT - Special Report, O'Reilly Media, Pages 15, Inc., 2009.
8. R. Cervenec, R. Burget: Identifying Expression of Emotions in Czech Text Using Semantic Relations for Dimension Reduction, Elektrorevue, Vol. 2 No. 3, Pages 16-21, 2011.
9. T. Mullen, N. Collier: Sentiment analysis using support vector machines with diverse information, National Institute of Informatics Tokyo, 2004.
10. R. Prabowo, M. Thelwall: Sentiment Analysis - A Combined Approach, School of Computing and Information Technology University of Wolverhampton, Pages 21, 2009.
11. Y. Wu, Q. Zhang, X. Huang, L. WuPhrase: Dependency Parsing for Opinion Mining, Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Pages 15331541, 2009.
12. P. Sanda: Determination of basic form of words, Brno University of Technology, Pages 68, 2011.
13. T. Hartmann, S. Klenk, A. Burkovski, G. Heidemann: Sentiment Detection with Character n-Grams, University of Stuttgart, Pages 5, 2008.
14. J. Kamps, M. Marx, R.J. Mokken, M. de Rijke: Using WordNet to measure semantic orientation of adjectives, Language and Inference Technology Group ILLC, University of Amsterdam, Pages 4, 2004.
15. C. Fellbaum : Wordnet: An Electronic Lexical Database. MIT Press, Cambridge, MA, Pages 423, 1998.

16. A. Krištofic and M. Bieliková: Improving adaptation in web-based educational hypermedia by means of knowledge discovery. In Proc. of HT 2005 Sixteenth ACM Conference on Hypertext and Hypermedia, ACM Press, Sept. 2005, Pages 184-192, 2005.
17. B. Pang, L. Lee: A Sentimental Education: Sentiment Analysis using Subjectivity Summarization based on minimum cuts. In Proceedings of the Association for Computational Linguistics (2004)
18. P. Šaloun, Z. Velart, J. Nekula: Towards automated navigation over multilingual content. Springer, Studies in Computational Intelligence 418, Pages 203-229, 2013
19. C. J. Van Rijsbergen: Information Retrieval (2nd ed.). Butterworth-Heinemann, Newton, MA, USA, Pages 147, 1979.